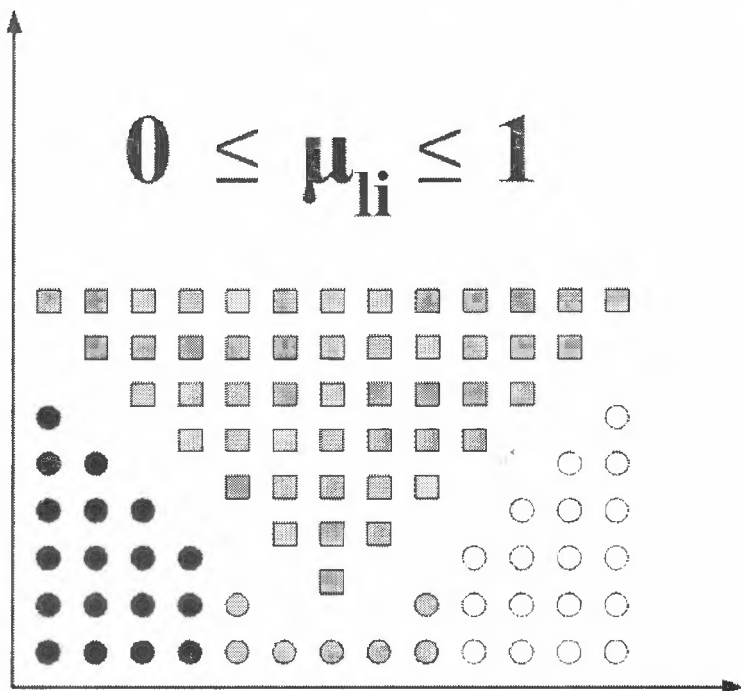


VISIT...

LANZAROTE
Caliente.COM

Д.А. Вятчин

НЕЧЕТКИЕ МЕТОДЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ



Д.А. Вятчин

Нечеткие методы автоматической классификации

Монография

Минск
УП «Технопринт»
2004

Рецензенты:

профессор кафедры математического моделирования
и анализа данных БГУ, доктор физ.-мат. наук Жук Е.Е.,
профессор кафедры управления войсками в бою
и операции Военно-инженерного университета
Министерства обороны Российской Федерации,
доктор техн. наук, полковник Сизов В.А.

Вятченин Д.А.

В 99 Нечеткие методы автоматической классификации: Монография /
Д.А. Вятченин –Мн.: УП «Технопринт», 2004 –219 с.
ISBN 985-464-529-0

Книга посвящена рассмотрению современного состояния нечетких методов автоматической классификации. Приводятся описания наиболее известных нечетких кластер-процедур и результатов вычислительных экспериментов.

Предназначена для научных и инженерно-технических работников, проводящих исследования в областях прикладной математики, информатики, кибернетики, робототехники и искусственного интеллекта; может быть полезной специалистам, проводящим исследования в областях управления, экономики, социологии, медицины, биологии, логики и методологии науки, а также студентам и аспирантам математических и инженерно-технических специальностей высших учебных заведений.

УДК 51-7

ББК 22.1

ПРЕДИСЛОВИЕ

Одним из наиболее интересных и многообещающих подходов к анализу многомерных явлений и процессов, как известно, являются методы автоматической классификации, именуемые также методами кластерного анализа или методами распознавания образов с самообучением. С момента возникновения кластерного анализа как раздела общей теории статистических решений, начиная с работ К. Чекановского, П. В. Терентьева, Р. С. Трайона и других исследователей, теория автоматической классификации непрерывно развивалась и совершенствовалась, в результате чего был разработан мощный математический аппарат и сотни разнообразных эффективных алгоритмов классификации в условиях отсутствия обучающих выборок. Проблемам разработки и применения методов автоматической классификации посвящены десятки тысяч статей и монографий зарубежных и отечественных исследователей, среди которых следует отметить ставшие классическими работы Р. О. Дуды и П. Е. Харта [79], К. С. Фу [88], К. Фукунаги [42], Дж. А. Хартигана [93], С. Ватанабэ [183], Ю. С. Харина [110], Е. Е. Жука [23], С. А. Айвазяна, В. М. Бухштабера, И. С. Енюкова и Л. Д. Мешалкина [37], И. Д. Манделя [31].

После публикации в 1965 году основополагающей статьи выдающегося американского математика Лофти А. Заде предложенная им теория нечетких, или, как иногда ее именуют в специальной литературе, размытых множеств получила бурное развитие, а предложенные ей методы нашли применение практически во всех областях человеческой деятельности, что, естественно, нашло свое отражение и в теории автоматической классификации.

Несмотря на богатые традиции отечественной школы нечетких систем, число книг на русском языке по теории нечетких множеств и различным сферам ее применения продолжает оставаться незначительным, а книги, посвященной рассмотрению нечетких методов автоматической классификации, нет. В то же время за рубежом наблюдается значительный рост интереса к данной области исследований, на что указывает появление ряда монографий, среди которых следует отметить работы Е. Бейкера [46], Дж. Беждека [59], С. Ф. Боклиша [64], Ф. Хеппнера, Ф. Клавонна и Р. Крузе [101], С. Мийамото [128]. Стремление заполнить означенный пробел, познакомить отечественных специалистов с достижениями нечеткого подхода в кластерном анализе и привлечь к исследованиям в этой области новых специали-

стов явилось стимулом к написанию и изданию предлагаемой вниманию читателя книги.

Глава 1 «Проблема неопределенности в задачах автоматической классификации» является вводной и рассматривает специфические особенности автоматической классификации как одного из подходов к выделению однородных групп объектов. Выделяются основные направления решения задач автоматической классификации, такие как эвристическое, иерархическое, оптимизационное и аппроксимационное, а также рассматривается содержательная постановка задачи автоматической классификации. Особое внимание уделяется рассмотрению методологических аспектов обработки различных видов неопределенности и логико-гносеологических аспектов применения теории нечетких множеств в задачах автоматической классификации.

Глава 2 «Основные понятия теории нечетких множеств» содержит изложение основных понятий и определений теории нечетких множеств.

Глава 3 «Методы нечеткого подхода к решению задач автоматической классификации» включает рассмотрение особенностей эвристических, оптимизационных и иерархических методов нечеткого подхода к решению задач автоматической классификации, а также детальное описание наиболее известных нечетких кластер-процедур.

Глава 4 «Сравнительный анализ нечетких методов автоматической классификации» посвящена рассмотрению общей схемы сравнения нечетких методов автоматической классификации и анализу роли сравнения различных нечетких кластер-процедур при обработке данных в прикладных исследованиях, а также проблемам оценки, представления и интерпретации результатов нечеткой классификации. Приводятся результаты вычислительных экспериментов, сравнительный анализ которых позволяет определить эффективность некоторых алгоритмов.

Глава 5 «Методологические аспекты нечеткого подхода к решению задачи автоматической классификации» рассматривает общую схему выбора конкретного нечеткого метода автоматической классификации в процессе проведения прикладного исследования, а также вопросы, связанные с обоснованием задаваемых параметров выбранной процедуры. Кратко освещаются некоторые особенности применения нечетких методов автоматической классификации при решении задач обработки изображений, и приводится обзор других нечетких методов распознавания образов.

Автор выражает огромную признательность глубокоуважаемым рецензентам: доктору физико-математических наук, профессору кафедры математического моделирования и анализа данных Белорусского государственного университета Евгению Евгеньевичу Жуку и доктору технических наук, профессору кафедры управления войсками в бою и операции Военно-инженерного университета Министерства обороны Российской Федерации, полковнику Валерию Александровичу Сизову за внимательное ознакомление с рукописью и сделанные ценные замечания. Особую благодарность автор выражает доценту кафедры интеллектуальных информационных технологий Белорусского государственного университета информатики и радиоэлектроники, кандидату технических наук Степановой Маргарите Дмитриевне за обсуждение многочисленных вопросов и полезные рекомендации в процессе подготовки книги, способствовавшие улучшению ее содержания, а также пионерам отечественной школы теории нечетких множеств: кандидату физико-математических наук, ведущему научному сотруднику Вычислительного центра Российской академии наук, член-корреспонденту Международной академии информатизации Аверкину Алексею Николаевичу, кандидату физико-математических наук, директору по научной работе севавтопольской компании ДАТА С Силону Валерию Болеславовичу и кандидату технических наук, доценту кафедры компьютерных систем автоматизации производства Московского государственного технического университета им. Н. Э. Баумана, члену-корреспонденту Международной академии информатизации Тарасову Валерию Борисовичу, одоббившим идею написания книги. Кроме того, автор благодарит заведующего кафедрой математического моделирования и анализа данных Белорусского государственного университета, доктора физико-математических наук, профессора Харина Юрия Семеновича, заведующего кафедрой информационных технологий автоматизированных систем Белорусского государственного университета информатики и радиоэлектроники, доктора технических наук, профессора Муху Владимира Степановича, заведующего кафедрой математического обеспечения автоматизированных систем управления Белорусского государственного университета, кандидата физико-математических наук, доцента Краснопрошина Виктора Владимировича, доцента кафедры математического обеспечения автоматизированных систем управления Белорусского государственного университета, кандидата физико-математических наук Образцова Владимира Алексеевича, заведующего отделом логики и ме-

тодологии научного познания Института философии Национальной академии наук Беларуси, академика Национальной академии наук Беларуси, доктора философских наук, профессора Широканова Дмитрия Ивановича, заведующего кафедрой статистики и экономической кибернетики Экономической академии им. Оскара Лангега во Вроцлаве, профессора Валентега Остасевича, заведующего отделом применения методов системного анализа Института системных исследований Польской академии наук, доктора Яна Овсиньского, вице-директора Института системных исследований Польской академии наук, действительного члена Польской академии наук, профессора Януша Каспшика, сотрудника отдела интеллектуальных информационных систем Института системных исследований Польской академии наук, профессора Эулалию Шмидт, руководителя Института статистики и теории вероятностей Венского технического университета, профессора Рейнхарда Фиртла за интерес к исследованиям, проводимым автором, и оказываемую всестороннюю помощь и поддержку. Автор выражает признательность Д. И. Новикову, А. О. Макарову, С. Г. Крупскому, А. П. Богодязу, В. В. Клемпачу, С. В. Полаженко, А. Л. Кондратенко, Е. А. Яркову, В. В. Каллауру, А. П. Слиборскому, Н. Ю. Чалому, Д. М. Добышеву и Ю. В. Барко за разработку и тестирование программного обеспечения, проведение вычислительных экспериментов и помощь в подготовке иллюстративного материала книги. Автор также считает своим долгом выразить благодарность редактору книги Галине Владимировне Ширкиной и коллективу УП «Технопринт» в лице директора Антона Петровича Аношко за кропотливый и добросовестный труд в процессе работы с рукописью и издание книги. За все допущенные ошибки и недостатки книги ответственность несет исключительно автор.

ПРОБЛЕМА НЕОПРЕДЕЛЕННОСТИ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

1.1. Понятие однородности и проблема классификации объектов

1.1.1. Основные подходы к решению проблемы выделения однородных групп объектов

Взаимосвязь понятий «однородность» и «классификация» представляется очевидной даже на интуитивном уровне. С. И. Ожегов в «Словаре русского языка» дает следующие трактовки этих понятий: «Однородный — относящийся к тому же роду, разряду; одинаковый» [36, с.382]; «классифицировать — распределить по группам, разрядам, классам» [36, с.238]. Если обе этих трактовки объединить в одну, то классификацию можно понимать как разбиение множества объектов на однородные группы или классы. В таком случае под однородностью подразумевается наличие у объектов одного класса общих свойств или признаков, определяющих некоторое сходство данных объектов и служащих основанием для отнесения этих объектов к одному классу.

Вместе с тем, во многих областях математики, к примеру, в прикладной статистике, понятие однородности оказывается основополагающим, так как обработка статистических данных производится только в однородных группах [22, с.11]. Требование однородности исследуемого множества объектов не ограничивается только лишь определением наблюдаемого объекта, так как любое реальное множество объектов являет собой систему дифференцированных, различающихся между собой элементов, что делает задачу разбиения исходного множества исследуемых объектов на однородные подмножества приоритетной при анализе систем любой природы: технических, биологических, социально-экономических.

Одной из основных особенностей задачи классификации является наличие как качественных, так и количественных признаков в описании объектов исходного множества, в силу чего при выделении однородных групп различают такие виды группировки исходных данных, как структурная и типологическая. *Структурной группировкой*

именуется разбиение качественно однородного исходного множества объектов на классы, которые характеризуют общее строение исходного множества объектов [40, с.96]. *Типологической группировкой* называется разбиение исходного множества объектов на классы определенного качества. Таким образом, структурная группировка представляет собой способ выделения количественно однородных групп объектов, а типологическая — способ выделения качественно однородных групп. При сопоставлении этих определений происходит своеобразное противопоставление категорий качества и количества, сильно упрощающее понимание этих категорий и являющееся, вообще говоря, неправомерным [31, с.7].

Если в основе типологической группировки находится некоторый качественный признак, причем единственный, то задача классификации, как правило, решается элементарно, однако в подавляющем большинстве случаев ее необходимо проводить по количественным признакам, что в значительной степени усложняет задачу. В таком случае на начальных этапах исследования рассуждения о качественной однородности исходных данных лишены всякого смысла, поскольку качественная однородность данных может быть установлена только в результате проведения анализа, основой которого, как справедливо указывали И. И. Елисеева и В. О. Рукавишников, должен быть синтез «теоретических концепций и опыта прошлых исследований» [22, с.15].

Таким образом, представляется нецелесообразным различать методы выделения качественно и количественно однородных групп, однако имеет смысл, как отмечал И. Д. Мандель, «говорить только о непрерывном синтезе этих категорий в процессе классификации» [31, с.10]. Методы выделения однородных групп объектов, в связи с вышеизложенным замечанием, условно объединяются в следующие основные подходы [31]:

Вероятностный подход основан на предположении о том, что объекты, принадлежащие одному из выделяемых классов, описываются одинаково распределенными случайными векторами, а для различных классов характерны различные распределения вероятностей. В специальной литературе этот подход традиционно именуется расщеплением смесей распределений, где каждый класс понимается как некоторая параметрически заданная одномодальная совокупность, а наблюдения над объектами, подлежащими классификации, трактуются как выборка из смеси таких совокупностей, так что задача заключа-

ется в разделении этих совокупностей, исходя из значений параметра, определяющего совокупность, и некоторых предположений, к примеру, о числе классов.

Вариативный подход состоит в разбиении множества объектов по выбранному исследователем признаку на интервалы группирования, в результате чего исходное множество объектов разбивается на группы таким образом, что объекты одной группы находятся на относительно небольшом расстоянии друг от друга. В многомерном же случае, при наличии нескольких признаков, данный подход представляет собой комбинационную группировку, для которой характерно поочередное использование признаков для выделения групп. Такой подход, когда единственный признак используется для разбиения всего множества объектов на группы, а также в случае поочередного использования различных признаков, когда каждый из них применяется для выделения одной группы, называется *монотетическим* [22, с.7].

Структурный подход базируется на представлении об объектах как точках в многомерном пространстве. В этом случае задача состоит в выделении из исходного множества многомерных точек однородных подмножеств таким образом, чтобы элементы каждого подмножества были в определенном смысле сходны между собой, а сами подмножества — классы объектов — отличались бы друг от друга, так что отыскивается своего рода «естественное» расслоение исходного множества на классы. Данный подход иногда именуется геометрическим, поскольку, используя понятия расстояния между объектами и расстояния между классами, выделяет геометрически удаленные группы. Наиболее последовательно геометрический подход реализован в методах кластерного анализа, которые в специальной литературе называются также методами автоматической классификации, численной таксономией или распознаванием образов с самообучением. В отличие от монотетического подхода к проблеме классификации объектов, кластерный анализ использует *одновременно все признаки* и называется *политетическим*.

Подробное исследование взаимосвязей между вышеизложенными подходами к решению проблемы выделения однородных групп объектов проведено И. Д. Манделем [31]. Вместе с тем, анализируя соотношение вероятностного и структурного подходов, в первую очередь необходимо отметить то обстоятельство, что многие зарубежные исследователи, такие как Дж. Хартиган [93], К. Фукунага [42], М. Вонг [187], рассматривают кластер-анализ чрезмерно широко,

включая в него и задачи расщепления смесей, то есть задачи классификации в условиях отсутствия обучающих выборок, когда исходные данные об исследуемых объектах имеют вероятностную природу и каждый класс интерпретируется как одномодальная генеральная совокупность при неизвестном значении определяющего ее параметра, а классифицируемые объекты рассматриваются как выборки из смеси таких генеральных совокупностей. Как способ представления исходных данных понятие смеси использует также известный польский исследователь Я. В. Овсиньски [134, с.392] при рассмотрении общей постановки задачи кластер-анализа. В отечественной литературе подобное рассмотрение автоматической классификации прослеживается в работах М. И. Шлезингера [44] и А. В. Миленького [32]. Е. Е. Жук и Ю. С. Харин [23, с.17-19] также указывают на существование в кластер-анализе вероятностного и геометрического подходов, отдавая предпочтение первому. Необходимо указать, что применимость методов расщепления смесей вероятностных распределений к решению задач классификации зависит от обоснованности предположений о вероятностной природе исходных данных и корректности выдвигаемой гипотезы о распределении вероятностей, описывающих классы объектов, тогда как успешное применение геометрических методов классификации зависит только от адекватности выбранной меры близости объектов. Отнесение же группы вероятностно-статистических методов классификации в условиях отсутствия обучающих выборок к кластерному анализу в силу причин методологического характера представляется спорным, так что следует говорить не о вероятностном подходе к решению задачи автоматической классификации, а о теоретико-вероятностной модификации задачи кластер-анализа, как это было предложено С. А. Айвазяном [37, с.146].

Касательно соотношения вариативного и структурного подходов, здесь лишь укажем, что при использовании вариативного подхода, являющегося разновидностью типологической группировки, классы имеют субъективный характер, а сама группировка является полностью управляемой, так что «естественное» расслоение, отыскиваемое методами кластерного анализа, в случае применения вариативных методов не имеет места. Главным же отличием двух подходов является то, что понятия «близость» и «сходство» объектов в типологической группировке неформализованы, в отличие от структурного подхода, где они формализованы и выражаются рядом соотношений. Достаточно полный обзор метрик и мер близости, формализующих

понятие сходства и используемых в задачах классификации, содержится в работах [29], [180].

1.1.2. Общая постановка задачи автоматической классификации и основные направления ее решения

Рассмотрим формы представления исходных данных в задачах классификации в условиях отсутствия обучающих выборок. Пусть $X = \{x_1, \dots, x_n\}$ — множество n объектов, каждый из которых характеризуется m признаками, так что каждый объект рассматривается как точка в m -мерном признаковом пространстве. Тогда исходные данные могут быть представлены матрицей

$$X_{m \times n} = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \dots & \dots & \dots & \dots \\ x_1^m & x_2^m & \dots & x_n^m \end{pmatrix}, \quad (1.1)$$

именуемой также матрицей «объект-свойство» [37, с.143], где x_i^t представляет собой значение t -го признака на i -м объекте. Таким образом, i -й столбец этой матрицы $X_i = (x_1^1, x_2^1, \dots, x_n^1)$ полностью характеризует объект x_i и будет интерпретироваться как точка в m -мерном признаковом пространстве $I^m(X)$.

Если же известны взаимные расстояния между объектами множества X , то исходные данные могут быть представлены в форме матрицы взаимных расстояний объектов $x_1, \dots, x_n$ или, что то же самое, точек $X_1, \dots, X_n$:

$$d_{n \times n} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}, \quad (1.2)$$

где величина d_{ij} представляет отдаленность объектов x_i и x_j друг от друга. Как вариант такой формы представления исходных данных может быть использована матрица

$$r_{n \times n} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}, \quad (1.3)$$

где величина r_{ij} , напротив, характеризует близость объектов x_i и x_j друг к другу. Очевидно, что чем более сходны объекты x_i и x_j , тем ближе расположены соответствующие точки в признаковом пространстве; таким образом, из неравенства $d_{ki} \geq d_{ij}$ должно следовать выполнение неравенства $r_{ki} \leq r_{ij}$. Поскольку переход от матрицы $d_{n \times n}$ к матрице $r_{n \times n}$ и наоборот, как правило, оказывается элементарным, то в дальнейших рассмотрениях вместо обозначений метрики d_{ij} или меры сходства r_{ij} будет использоваться одно общее обозначение функции близости или отдаленности ρ_{ij} , а вместо матриц $d_{n \times n}$ и $r_{n \times n}$ — соответственно матрица

$$\rho_{n \times n} = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & \rho_{nn} \end{pmatrix}, \quad (1.4)$$

но при необходимости будет осуществляться возвращение к рассмотрению величины расстояния d_{ij} или величины близости r_{ij} и матриц $d_{n \times n}$ и $r_{n \times n}$ соответственно, что будет оговариваться особо. В специальной литературе матрица $\rho_{n \times n}$ любого типа носит название матрицы «объект-объект» [37, с.143].

В наиболее общем виде проблема классификации объектов в условиях отсутствия обучающих выборок состоит в разбиении на заранее известное либо нет число однородных, в определенном смысле, классов всего исходного множества объектов $X = \{x_1, \dots, x_n\}$, представленного в виде матрицы «объект-свойство» или матрицы «объект-объект».

Следует сразу оговорить, что при такой постановке задачи исходные данные могут иметь и вероятностную природу, однако поскольку объектом рассмотрения является кластерный анализ, то следует придерживаться взгляда о геометрической природе исходных данных.

Необходимо также отметить, что аналогичным образом интерпретируется исходная информация в задаче классификации совокупности признаков, характеризующих множество объектов, представленных в виде матрицы «объект-свойство». Разница заключается в том, что каждый объект множества $X = \{x_1, \dots, x_n\}$ представлен соответствующим столбцом матрицы $X_{m \times n}$, тогда как признак будет задаваться соответствующей строкой матрицы $X_{m \times n}$. Поскольку постановка задачи и основная методологическая схема исследования и для задачи классификации объектов, и для задачи классификации признаков являются общими [37, с.144-145], то дальнейшее рассмотрение проблемы классификации будет проводиться применительно к объектам, а в случае задачи классификации признаков этот момент будет оговариваться особо.

При интерпретации объектов как точек в соответствующем пространстве признаков возникает следующая задача: *разделить совокупность точек $\{X_i\}, i = \overline{1, n}$ в пространстве $I^m(X)$ на однородные классы таким образом, чтобы точки, принадлежащие одному классу, находились бы относительно близко друг от друга, а сами классы различались бы между собой. Полученные в результате разбиения группы именовются кластерами, классами, образами или таксонами, а методы их обнаружения соответственно называются кластерным анализом, автоматической классификацией, распознаванием образов с самообучением или численной таксономией.* Следует указать, что термин «кластер» (cluster), имеющий английское происхождение, понимается в математике как группа элементов, характеризуемых некоторым общим свойством, а термин «таксон» (taxon), используемый в биологии и также заимствованный из английского языка, в задаче кластерного анализа понимается как систематизированная группа любой категории.

При такой постановке задачи оказывается необходимым формальное определение понятий однородности и близости объектов.

Матрица $\rho_{m \times n}$ однозначно задает геометрическую структуру множества объектов $X = \{x_1, \dots, x_n\}$, однако координаты соответствующих точек оказываются неизвестными. Если же исходные данные заданы матрицей $X_{m \times n}$, то точки $\{X_i\}, i = \overline{1, n}$ представляют собой, как указывалось выше, векторы-столбцы геометрических координат наблюдений в m -мерном признаковом пространстве $I^m(X)$. Задавая

способ вычисления величины ρ_{ij} , характеризующей отдаленность или близость объектов x_i и x_j исходного множества X , однородность можно определить следующим образом: близкие в смысле функции $\rho(x_i, x_j), x_i, x_j \in X, i, j = \overline{1, n}$ объекты считаются однородными и попадают в один класс. Близость объектов исходного множества, в свою очередь, определяется сопоставлением значения функции $\rho(x_i, x_j), x_i, x_j \in X, i, j = \overline{1, n}$ для каждой пары объектов с некоторым пороговым значением. Способ же вычисления величины ρ_{ij} можно задать только в случае, когда известны все координаты точек $\{X_i\}, i = \overline{1, n}$. Если же способ вычисления ρ_{ij} задан, то возможен переход от матрицы $X_{m \times n}$ к матрице $\rho_{n \times n}$. Обратный переход от матрицы $\rho_{n \times n}$ к матрице $X_{m \times n}$ осуществляется с помощью аппарата многомерного метрического шкалирования. Проблема выбора метрики или меры близости детально рассматривается С. А. Айвазяном [37, с.147-153] и И. Д. Манделем [31, с.26-35].

Если алгоритм разбиения множества объектов задан, то проблема выбора функции ρ_{ij} выступает на первый план, поскольку от этого полностью зависит разбиение множества объектов на классы. Выбор меры близости или расстояния определяется, во-первых, конечной целью проводимого исследования, природой исходных данных, а также некоторыми предположениями, к примеру о форме кластеров и их взаимном расположении. Таким образом, способ вычисления величины ρ_{ij} должен быть адекватен геометрической структуре исходного множества объектов; в этой связи С. А. Айвазян совершенно справедливо подчеркивал, что «выбор метрики (или меры близости) является узловым моментом исследования» [37, с.147].

Для анализа структуры множества объектов, а также при разработке различных алгоритмов автоматической классификации иногда оказывается необходимым рассматривать близость не только отдельных объектов, но и целых классов, для чего также используются различные расстояния и меры близости, что более подробно рассматривается С. А. Айвазяном в работе [37, с.153-156].

Если в процессе исследования имеются некоторые предположения о свойствах кластеров или принципах объединения объектов в классы, то оказывается вполне естественным формально определить понятие кластера и построить процедуру, выделяющую из исходного

множества объектов группы, удовлетворяющие данному определению. Подобный подход является основой **эвристического направления** решения задачи кластерного анализа, а алгоритмы разбиения исходного множества объектов на кластеры с заранее заданными свойствами, соответствующие группе методов этого направления, носят общее название *эвристических алгоритмов*.

Если же целью исследования является не разбиение исходного множества объектов на классы, а выявление стратификационной структуры исходного множества, то применяются методы, объединенные в **иерархическое направление** кластер-анализа. Группа иерархических методов объединяет *агломеративные и дивизимные алгоритмы* кластер-анализа. Для алгоритмов, соответствующих методам агломеративного подхода иерархического направления, предполагается вначале, что каждый объект представляет собой отдельный класс, после чего элементы и их группы объединяются, пока в результате последовательного объединения не получится исходное множество. Алгоритмы же дивизимной группы методов, напротив, исходят из предположения об исходном множестве как одном классе, который постепенно делится на все более мелкие, вплоть до отдельных элементов, группы объектов. Результаты работы иерархических алгоритмов обычно представляются в виде дендрограммы или графа иерархии.

Очевидно, что одно и то же множество объектов можно разбить на кластеры различными способами или при использовании одного метода можно получить целую группу различных разбиений. В таком случае имеет смысл определить качество разбиений с целью выбора наилучшего разбиения, то есть сформулировать количественный критерий, в соответствии с которым можно было бы предпочесть одно разбиение другому. Для формулировки представлений о качестве классификации в постановку задачи вводится функционал качества разбиения, задающий способ сопоставления с каждым разбиением P числа $Q(P)$, которое оценивает в некоторой шкале степень оптимальности разбиения P . То разбиение P^* , на котором выбранный функционал достигает экстремального значения, считается наиболее предпочтительным. Данное направление в кластерном анализе называется **оптимизационным** и вводит задачу автоматической классификации в сугубо математическое русло, так что в математической постановке задача автоматической классификации сводится к поиску оптимального разбиения P^* и формулируется в виде

$$Q(P) \rightarrow_{\text{PeII}} \text{extr}, \quad (1.5)$$

где Π — множество всех возможных разбиений исходного множества объектов X . Математические методы и реализующие их кластер-процедуры, доставляющие экстремум выбранному функционалу, соответственно называют *оптимизационными*.

В продолжение рассмотрения оптимизационных методов решения задачи автоматической классификации следует указать на то обстоятельство, что, как отмечал С. А. Айвазян, «в статистической практике выбор функционала качества разбиения $Q(P)$ обычно осуществляется весьма произвольно, опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо точную формализованную схему... Однако ряд распространенных в статистической практике функционалов качества удастся постфактум обосновать и осмыслить в рамках строгих математических моделей» [37, с.181].

Если классификация, которую требуется найти, описывается матрицей определенной структуры, к примеру, матрицей отношения эквивалентности, то задача заключается в оценке параметров искомой структуры так, чтобы искомая структура минимально отличалась бы от исходной структуры. Иными словами, отношение, отвечающее исходным данным, необходимо аппроксимировать отношением, которое отвечает представлению о наилучшей классификации, так что это направление решения задач кластер-анализа именуется **аппроксимационным**; в этом случае проблема сводится к следующей оптимизационной задаче:

$$\|M\| \rightarrow \text{extr}, \quad (1.6)$$

где $M = J - BX$. В этом равенстве J обозначает отношение, которое требуется найти, X — матрица исходных данных, B — оператор перехода от X к J , а $\|M\|$ обозначает некоторую норму. На практике применяют методы аппроксимации как матрицы «объект-свойство», так и матрицы «объект-объект».

Рассмотренные выше группы методов решения задач кластер-анализа не являются строгими и не исчерпывают всего богатства методов классификации объектов в условиях отсутствия обучающих выборок; многие алгоритмы находятся на стыке вышеизложенных направлений. Проблема же классификации собственно алгоритмов кластер-анализа была обозначена около тридцати лет назад и остается актуальной по настоящее время [31, с.39].

1.2. Подход к решению задачи автоматической классификации с позиции теории нечетких множеств

1.2.1. Основные концепции неопределенности в задачах автоматической классификации

Собственно процесс решения задачи классификации, независимо от природы исходных данных, в общем, состоит из восьми этапов [37, с.42-43]: установочного, на котором формулируется постановка задачи на содержательном уровне; постановочного, в ходе которого определяется тип прикладной задачи в терминах теории классификации; информационного, состоящего в выработке плана сбора исходной информации, ее предварительном анализе и редактировании; априорного математико-постановочного, заключающегося в выборе на основании выводов, полученных в результате реализации предыдущих этапов, базовых математических моделей для математической постановки конкретной задачи классификации; разведочного, предусматривающего применение специальных методов предварительного анализа исходных данных с целью выявления их вероятностной и геометрической природы; апостериорного математико-постановочного, в процессе которого уточняется выбор базовой математической модели с учетом результатов реализации разведочного этапа процесса решения задачи классификации; вычислительного, целью которого является программная реализация выбранного математического аппарата для решения конкретной задачи; итогового, на котором производится анализ и интерпретация результатов проведенного исследования. Таким образом, вид задачи классификации определяется в результате реализации первых трех этапов процесса исследования; к примеру, если предварительная выборочная информация отсутствует, а априорные сведения о классах объектов являют собой лишь некоторые предположения самого общего характера, то задача относится к классу задач распознавания образов с самообучением.

Вместе с тем, на практике зачастую оказывается, что задаче свойственна нечеткость [14], значительно затрудняющая или вообще делающая невозможным получение решения, так что на первый план выходит проблема устранения нечеткости, присущей задаче классификации. Понятие нечеткости является общенаучным и может быть определено как внешнее выражение качества внутренней основы яв-

лений, специфика которого заключается в непрерывности перехода от отсутствия проявления к полному выявлению качества предметов, свойств и отношений реального мира, что находит свое отражение в познавательной и мыслительной деятельности индивида. Содержание понятия нечеткости включает в себя последовательный ряд абстракций более низкого уровня. По отношению к человеческому сознанию выделяются такие категориальные виды нечеткости, как объективная и субъективная нечеткость; в свою очередь, объективная нечеткость может характеризоваться как стохастической, так и нестохастической детерминированностью. Объективная стохастическая нечеткость имеет такие формы проявления, как неопределенность и случайность. В данном случае неопределенность выступает в качестве нечеткой закономерности проявления свойств предмета, а случайность может быть определена как событие, имеющее нечеткое основание. Формами проявления объективной нестохастической нечеткости являются недетерминированность, рассматриваемая как нечеткость связи между предметами, свойствами или отношениями; размытость, характеризующая границы явлений, процессов, предметов, а также их классов и, кроме того, имена и область применимости предиката в логике; неоднозначность, определяемая как нечеткость значения признака объекта; неполнота, представляющая собой отсутствие всей возможной информации о рассматриваемом предмете или явлении, частными случаями которой выступают недостаточность как отсутствие необходимой информации и неадекватность как описание предмета по аналогии с рассмотренными ранее; неточность, являющаяся нечеткостью измерения или вычисления; неопределенность, трактуемая как нечеткость предела проявления характеристики предмета; случайность, определяемая как нечеткая реализация одной из нескольких существующих возможностей. Субъективная нечеткость имеет такие формы проявления, как неясность, под которой подразумевается нечеткость восприятия; размытость, которая в данном случае являет собой характеристику представления индивида о явлениях, процессах, предметах, свойствах, отношениях; недетерминированность, определяемая как свойство процесса логического вывода, производимого индивидом, в нечетких условиях; неоднозначность, понимаемая как нечеткость результата процесса интерпретации информации; неточность, которая в данном случае трактуется как мера соответствия знаний индивида о предмете объективным характеристикам рассматриваемого предмета; неопределенность, понимаемая как нечеткость отношения между объ-

ектом реального мира и представлением о нем в сознании индивида, а также как нечеткость смыслового значения имени, выражающего некоторое понятие.

Таким образом, виды и формы нечеткости могут быть упорядочены в соответствии с иерархией, представленной на рис. 1.1.

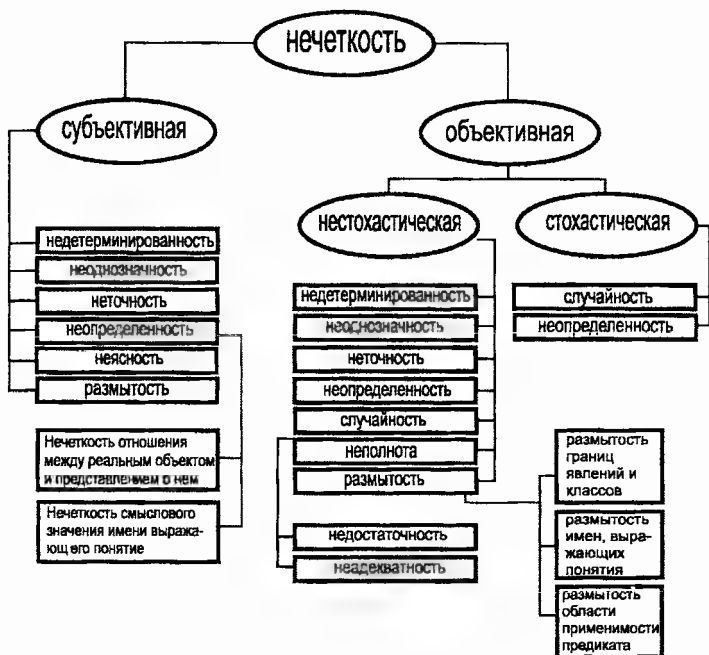


Рис. 1.1. Виды и формы нечеткости

Помимо рассмотренных форм нечеткости иногда выделяются и другие формы, такие, к примеру, как неизвестность и недоопределенность [35, с.8]; общим же для всех форм проявления нечеткости является то, что они характеризуют как свойства объектов, собственно свойств, отношений, процессов, явлений, понятий, так и особенности мыслительных и когнитивных процессов, присущих человеку.

В процессе анализа данных, в частности, в контексте задачи автоматической классификации, понятие нечеткости принимает вполне определенный смысл и оказывается характеристикой классов объектов, чему более соответствует понятие размытости, так что при анализе форм нечеткости, свойственной процессу анализа данных, более уместным представляется использование понятия неопределенности, под которой следует понимать нечеткость соответствия математической постановки задачи анализа данных предметно-содержательной установке на цели исследования и исходным данным. Поскольку постановка задачи на предметно-содержательном уровне включает формулировку целей исследования, определяющих тип задачи, выявление характера исходных данных и определение характера результатов исследования [37, с.42], то понятие неопределенности в той или иной степени характеризует каждую из этих составляющих.

Рассматривая процесс анализа данных в виде схемы, изображенной на рис. 1.2 [133, с.274], Я. В. Овсиньски и С. Задрожны отмечают, что «неопределенность может проникнуть в процесс анализа данных почти на каждом этапе... Эта неопределенность может не иметь вероятностного характера, что в большой степени свойственно природе переменных и процедуре их кодирования. Другой источник такой неопределенности относится к интерпретации содержания частных переменных, в случае, когда данные поступают от различных лиц» [133, с.273].

Процесс анализа данных в условиях нестохастической неопределенности, как правило, основан на ряде довольно жестких, не допускающих неопределенности предположений. Таким образом, традиционные методы анализа данных целесообразно применять только в тех случаях, когда подобные предположения достаточно обоснованы как с содержательной, так и с формальной точек зрения. С другой стороны, такое обоснование оказывается возможным в очень немногочисленных случаях, так что целесообразным оказывается «представить различные типы неопределенности» [133, с.275] для последующей обработки адекватными методами.



Рис. 1.2. Общая схема процесса анализа данных

Следовательно, одной из важнейших, с методологической точки зрения, задач в процессе анализа данных является установление форм проявления неопределенности и выбор соответствующих методологий для их обработки.

Поскольку объектом предпринятого рассмотрения является такой подход к анализу данных, как автоматическая классификация, то дальнейшее рассмотрение видов и форм неопределенности следует проводить, соответственно, применительно к задачам кластерного анализа и основой подобного рассмотрения могут послужить результаты, представленные в работе [19]. Конечными целями решения задачи автоматической классификации являются либо выделение четко выраженных классов статистически обследованных объектов в анализируемом многомерном пространстве, либо получение наглядного представления о стратификационной структуре классифицируемой совокупности объектов, либо оценка параметров структуры искомой классификации, минимально отличающейся от структуры исходных данных, так что в качестве аппарата решения задачи распознавания образов с самообучением выбираются либо алгоритмы оптимизационного или эвристического направлений, либо иерархические алгоритмы, либо алгоритмы, соответствующие аппроксимационному подходу в численной таксономии. Таким образом, неопределенность установок на главные цели исследования может повлечь некорректность математической постановки задачи автоматической классификации.

В процессе формулировки целей исследования неопределенность находит свое выражение в форме свойства интерпретации исследователем информации о предмете исследования и об исследуемой совокупности объектов ее реальным свойствам, то есть неясности, в форме характеристики соответствия знаний исследователя о свойствах совокупности объектов ее реальным свойствам, то есть неточности. Необходимо отметить, что обе формы неопределенности в данном случае носят субъективный характер.

Касательно характера результатов прикладного исследования, который также определяется на первых двух этапах решения задачи классификации, следует указать, что главным свойством процесса формулировки требований к результатам является неопределенность, носящая объективный характер. Как качественная характеристика, неопределенность выступает в процессе выявления природы искомой классификации в том смысле, должна ли полученная классификация являться размытой, или, пользуясь устоявшейся в специальной лите-

ратуре терминологией, нечеткой, или нет. Если же речь идет о форме и взаимном расположении кластеров, а также об их числе, то неопределенность выступает как количественная характеристика.

Если в качестве цели исследования выдвигается требование обнаружения «естественного расслоения» совокупности объектов на классы, то их число априорно оказывается неизвестным. Противоположной является ситуация, когда задача состоит в разбиении совокупности объектов на заранее известное число однородных классов. Число классов в таком случае выступает также в качестве исходных данных при постановке задачи классификации. Данные ситуации обуславливают выбор алгоритма классификации, поскольку во многих алгоритмах число кластеров является задаваемым параметром [31, с.156-158], [37, с.157-159]; к примеру, если выбрано оптимизационное направление решения задачи, то известность или неизвестность числа кластеров, на которые требуется разбить совокупность объектов, диктует тот или иной вид функционала качества разбиения. Гораздо сложнее обстоит ситуация, когда число кластеров нельзя указать однозначно. В этом случае неопределенность как характеристика всей задачи кластер-анализа проявляется в форме неоднозначности, которая, в свою очередь, оказывается следствием неточности как характеристики знаний исследователя о свойствах совокупности объектов и является субъективным типом неопределенности. В свою очередь, неясность в вопросе о природе искомой классификации влечет некорректный выбор метода классификации; так, если классификация носит размытый характер, но при этом алгоритм кластер-анализа не является нечетким, то полученные результаты также окажутся некорректными.

Неопределенность исходных данных в задачах численной таксономии также проявляется в различных формах. Первая из них, уже упоминавшаяся выше неоднозначность числа классов объектов исследуемой совокупности, может также рассматриваться как частный случай неполноты. Неоднозначность может также характеризовать число объектов совокупности, которую требуется разбить на классы.

Если исходные данные представлены в виде матрицы «объект-свойство» и значения некоторых признаков могут изменяться в зависимости от состояния объекта, примером чего может послужить значение звездной величины переменных звезд [21, с.206-207], представляя собой конечную совокупность либо интервал значений, неоднозначность выступает в качестве неопределенности значения признака

объекта; при переходе от матрицы вида «объект-свойство» к матрице вида «объект-объект» неоднозначность выступает в качестве неопределенности расстояния либо меры близости между объектами.

Поскольку значения признаков объектов могут быть измерены, вычислены либо определены путем экспертной оценки с некоторой погрешностью, то неопределенность исходных данных задачи численной таксономии в этом случае проявляется в форме неточности как свойства процесса измерения или вычисления. Неточность исходных данных также может быть характерна и для их представления в форме матрицы вида «объект-объект». Здесь также уместно отметить то обстоятельство, что неточность исходных данных, свойственная процессу измерения значений признаков объектов может возрасти при переходе от матрицы «объект-свойство» к матрице «объект-объект», поскольку вычисление расстояния между парой объектов, посредством которого осуществляется такой переход, сопровождается, как правило, погрешностью округления.

Одним из наиболее часто встречающихся типов неопределенности, свойственной исходным данным, является недостаточность, проявляющаяся в отсутствии части значений в матрице исходных данных. Данная форма неопределенности может быть характерна как для матриц вида «объект-свойство», так и для матриц вида «объект-объект», хотя в специальной литературе по анализу данных преимущественно рассматриваются матрицы первого вида. Недостаточность данных в специальной литературе именуется неполнотой, а отсутствие значений в матрице исходных данных называется пропусками в данных [30]. Примерами причин, вызывающих пропуски в данных, могут послужить ситуации, когда в процессе социологического опроса часть респондентов отказывается сообщить размер дохода, когда в промышленном производстве вследствие поломок оборудования не удастся получить данные для исследования, а также когда в период избирательной компании при опросе общественного мнения какая-то часть респондентов не окажет предпочтения одному кандидату перед другим. Пропускам в данных из первых двух примеров соответствуют истинные значения, которые были бы получены при более современном методе исследования или более высоком качестве промышленного оборудования, так что ненаблюдаемые значения естественно рассматривать как пропуски в данных. В третьем примере представляется менее правдоподобным, что за отсутствием ответа кроется предпочтение

определенному кандидату, так что рассматривать отсутствующие значения как пропуски менее естественно [30, с.11].

Описанные виды и формы неопределенности в задачах автоматической классификации не исчерпывают всего многообразия форм ее проявления и лишь очерчивают концептуальную рамку их рассмотрения, так что основные концепции неопределенности в задачах кластерного анализа могут быть представлены в виде схемы, изображенной на рис. 1.3.

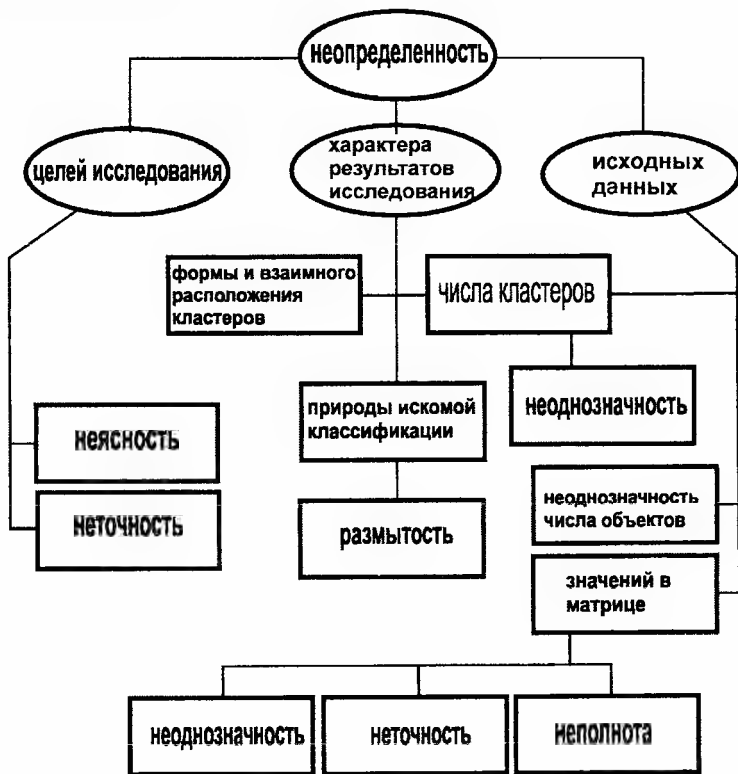


Рис. 1.3. Концепции неопределенности в задачах автоматической классификации

Каждой из рассмотренных форм неопределенности в задачах автоматической классификации соответствует определенная методология ее формализации или устранения. Касательно неопределенности целей исследования следует указать, что поскольку формы ее проявления носят субъективный характер, необходимым условием для ее устранения является полное описание предмета исследования, сбор всей возможной информации об объектах исследуемой совокупности, а также о свойствах самой совокупности. Кроме того, необходимым также является привлечение специалистов из той предметной области, к которой относится решаемая задача, не только к выработке предметно-содержательной установки на цели исследования и формирования этих целей в терминах анализа данных, но и к выработке плана сбора исходной информации, а также ее аттестации.

Для формализации или устранения неопределенности в процессе определения характера результатов исследования, являющейся, как правило, следствием отсутствия априорной информации о механизме порождения анализируемых данных, применяется аппарат разведочного анализа данных [37, с.473-487], основной целью которого, в свою очередь, является построение некоторой модели данных. Подобная модель данных, в общем случае, требует верификации. Разведочный анализ данных может применяться как к данным, заданным в форме матрицы «объект-признак», так и к данным, заданным в форме матрицы «объект-объект», причем если исходные данные имеют форму матрицы «объект-признак», то переменные могут быть измерены в различных шкалах. Главным элементом разведочного анализа данных является визуализация данных, предполагающая получение графического отображения исследуемой совокупности точек в многомерном пространстве, так что путем непосредственного визуального анализа изображения структуры исследуемой совокупности оказывается возможным определить форму и взаимное расположение кластеров, их число, а также природу искомой классификации. В случае пересечения кластеров, их соприкосновения или соединения цепочкой, а также когда исходные данные представляют собой неоднородное облако, искомая классификация оказывается размытой, так что при выборе математического аппарата для решения задачи кластер-анализа предпочтение следует отдать нечетким кластер-процедурам. Более того, если скопление точек имеет четко выраженную форму, имеет смысл обратиться к эвристическим алгоритмам, реализующим одно или несколько определений кластера [31, с.37]; последнее обстоятельство

отражает ситуацию, когда различные скопления точек имеют различную форму. Помимо применения аппарата разведочного анализа данных, с целью определения характера классификации, которую необходимо получить, требуется также минимизировать неопределенность исходных данных, для чего следует выяснить ее причины.

Если число кластеров, на которое требуется разбить исследуемую совокупность объектов, однозначно определить не удастся, имеет смысл обратиться к оптимизационному направлению решения задачи автоматической классификации и выбрать функционал качества разбиения с неизвестным числом классов. В случае, когда неоднозначность является объективной характеристикой числа объектов исследуемой совокупности, следует прибегнуть к аппарату нечетких интервалов, в частности, к аппарату треугольных нечетких чисел (L-R)-типа [34, с.105-106]. Нелишним будет отметить, что формализация или устранение неопределенности в данном случае оказывается необходимым условием для представления исходных данных в виде матрицы любой из указанных форм. В любом случае неоднозначность, являющаяся объективной характеристикой числа объектов исследуемой совокупности, влечет неполноту данных.

Наиболее важным для рассмотрения случаев неопределенности исходных данных в задачах автоматической классификации является неопределенность значений в матрице исходных данных. В случае численной неточности значений, имеющей объективный характер, целесообразно прибегнуть к другому аппарату или инструментарию для повышения точности измерения. С целью устранения неоднозначности значений переменных в матрице исходных данных можно применять метод, именуемый заполнением средними, когда вместо множества либо интервала значений переменной подставляется некоторое среднее значение этого множества или интервал значений. Возможно также применение метода, основанного на оценке с помощью регрессии множества значений переменной. Неполные данные, как правило, обрабатываются с помощью четырех групп методов: методов исключения некомплектных объектов, методов с заполнением, методов взвешивания и методов, основанных на моделировании [30, с.14-15].

Таким образом, методологии обработки основных видов объективной неопределенности в задачах автоматической классификации можно представить в виде таблицы 1.1.

Методологии обработки неопределенностей в задачах автоматической классификации

Неопределенность задачи автоматической классификации		Методология обработки неопределенности
Вид неопределенности	Форма проявления неопределенности	
Неопределенность характера результатов исследования	Неопределенность формы и взаимного расположения кластеров	Разведочный анализ данных
	Размытость природы искомой классификации	Нечеткие методы автоматической классификации
	Неоднозначность числа кластеров	Оптимизационные алгоритмы с функционалом качества при неизвестном числе кластеров
Неопределенность исходных данных	Неоднозначность числа объектов исходной совокупности	Нечеткие числа (L-R)-типа
	Неоднозначность значений переменных в матрице	Заполнение средними; заполнение с помощью регрессии
	Неточность значений переменных в матрице исходных данных	Инструментарий с высокой точностью измерения или вычисления
	Неполнота исходных данных	Методы исключения некомплектных объектов; методы с заполнением; методы взвешивания; методы, основанные на моделировании

Методологии обработки различных форм неопределенности, возникающей в задачах распознавания образов с самообучением, рассмотренные выше, не являются единственно возможными и не претендуют на статус универсальных; более того, возможными являются ситуации, когда в задаче автоматической классификации встречается несколько форм неопределенности. В таком случае оказывается целесообразным применение комбинированных методов либо поэтапное устранение неопределенности. Однако в любом случае первым этапом решения задачи кластер-анализа в условиях неопределенности должен являться детальный анализ задачи с целью выявления видов и форм

неопределенности, механизма их возникновения и взаимной обусловленности. Только после успешной реализации этого этапа исследования следует определять методологию обработки неопределенности, после чего производить обработку в соответствии с выбранной методологией. После формализации или устранения неопределенности следует произвести оценку результатов обработки и только в случае получения удовлетворительных результатов переходить к решению задачи классификации.

1.2.2. Гносеологические аспекты подхода к решению задач классификации с позиции теории нечетких множеств

Кластер-анализ представляет собой структурный подход к решению проблемы группировки многомерных объектов, основа которого заключается в представлении результатов отдельных наблюдений точками геометрического пространства с последующим выделением групп как «сгустков» этих точек, именуемых кластерами. Большинство существующих на сегодняшний день подходов и методов к решению задач кластер-анализа имеют своей основой эвристические соображения, которые возникают из конкретных приложений понятия классификации к некоторым частным классам однотипных задач, так что применение различных методов для конкретной задачи приводит к различным классификациям. Таким образом, механическое расширение области применимости частных процедур приводит к неудовлетворительным результатам и не позволяет постичь сущность причин, вызывающих таксономические различия. Зачастую это оказывается следствием различающихся между собой определений кластера, основывающихся на интуитивном представлении о кластере как о множестве объектов, подобных друг другу и отличных от объектов, не принадлежащих этому множеству. Традиционно формулируются требования, согласно которым кластер должен представлять собой непустое подмножество объектов, разные кластеры должны отличаться друг от друга и все объекты исходного множества должны быть расклассифицированы. В случае детерминистской постановки задачи последние два требования формулируются более жестко: оказывается необходимой дизъюнктивность кластеров и образы кластеров не должны пересекаться, то есть каждый объект исходного множества должен принадлежать только одному кластеру.

Подобные требования, в частности, требование об однозначности классификации, оказываются чрезмерно жесткими при анализе

сложных динамических систем, к которым можно отнести социально-экономические, биологические и иные виды систем, в которых центральное место занимают живые объекты, к примеру, человек. Такие системы профессор Л. А. Заде предложил именовать гуманистическими [24]. При подобного рода исследованиях возможность формулировки точных и в то же время содержательно значимых высказываний сводится к минимуму, за которым точность и релевантность становятся взаимоисключающими характеристиками подобных высказываний. Практика показывает, что традиционные статистические методы автоматической классификации зачастую не дают устойчивых результатов в случае, когда два или более кластеров соединяются цепочкой из внутренне связанных объектов выборки, когда кластеры имеют несферическую форму, когда кластеры являются линейно-несепарабельными множествами либо когда различаются плотности или объемы кластеров.

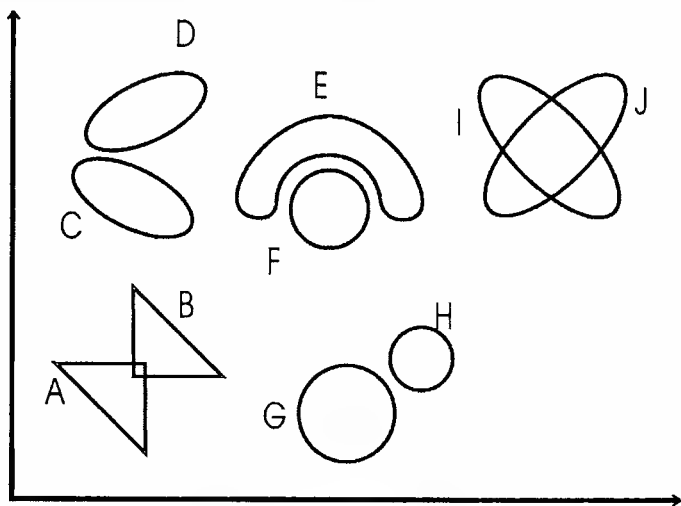


Рис. 1.4. Различные формы кластеров

На рис. 1.4 первый случай иллюстрируется кластерами А и В, второй случай представлен кластерами С и D, кластер Е демонстрирует третий из описанных выше случаев, а кластеры G и H — четвертый.

Еще более сложным оказывается случай пересечения кластеров, который иллюстрируется кластерами I и J.

Подобные примеры наглядно демонстрируют, что многие реальные системы обладают структурой, описание которой в рамках традиционных математических формализмов оказывается невозможным, так что человеческие суждения о поведении или состоянии подобных систем в действительности относятся не к какому-либо конкретному состоянию системы, а к совокупности различных состояний системы, границы между которыми оказываются объективно размыты, что, в свою очередь, можно продемонстрировать на следующем примере.

Предположим, что основой для классификации группы объектов выбран такой признак, как цвет объекта. В терминах разговорного языка описание значений переменной «цвет» является неточным, в отличие от численного представления диапазона длин волн, присущих каждому цвету. В процессе различения цветов человек основывается не на эталонном, а на собственном восприятии того или иного цвета, а этот субъективный диапазон длин волн, как показывают физиологические исследования, может колебаться в весьма значительных пределах. Таким образом, цвет, одним человеком воспринимаемый как «оранжевый», другим человеком воспринимается как «желтый», что с медицинской точки зрения оказывается вполне нормальным, поскольку патологией, в специальной литературе именуемой дальтонизмом, является неспособность человека различать три основных цвета — красный, зеленый и синий. Данный пример демонстрирует, что границы интервала длин волн, в естественном языке обозначаемого одним и тем же словом, для разных людей оказываются различными, то есть и множество волн, обозначаемое этим термином, оказывается размытым.

Вместе с тем, нечеткость может быть не только субъективной, но и объективной характеристикой. Нечеткость как характеристика собственно объектов является следствием многообразия признаков, характеризующих объект, а также динамики их изменения вследствие изменения структуры объекта. Примерами таких объектов могут служить, хамелеон — в зоологии, переменные звезды — в астрономии, самолет с изменяемой геометрией крыла — в технике. Касательно последнего примера можно указать, что для человека, не являющегося специалистом в области авиационной техники, один и тот же самолет, но при различных углах стреловидности передней кромки крыла, со-

ответствующих различным режимам полета, может быть воспринят как два совершенно разных типа самолетов, и если в качестве основы для классификации выбрать такие признаки, как размах крыла или угол стреловидности передней кромки крыла, то один и тот же самолет, взятый в различных ситуациях, попадет в различные кластеры.

Подобные примеры подтверждают тезис, выдвинутый профессором Л. А. Заде, что «большинство реальных кластеров размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен» [25, с.208]. Теория нечетких множеств, вводя понятие взвешенной принадлежности объекта множеству, предлагает гибкий аппарат для формального описания подобного рода ситуаций, так что вопрос заключается не в том, принадлежит ли объект классу, а в том, какова степень принадлежности объекта тому или иному классу.

Здесь целесообразно кратко рассмотреть логический аспект проблемы. Если в случае обычной, или, как еще именуют в специальной литературе, четкой классификации, объекты, находящиеся в области пересечения кластеров, оказываются однозначно принадлежащими обоим кластерам в равной степени, что схематически изображено на рис. 1.5, то в случае нечеткого подхода к решению задачи автоматической классификации кластеры представляют собой нечеткие множества, так что о принадлежности объекта тому или иному кластеру можно судить по его функции принадлежности, которая в данном случае будет выражать степень сходства объекта с типичным элементом кластера, что изображено, соответственно, на рис. 1.6.

Таким образом, нечеткий подход к решению задачи классификации в ряде случаев позволяет разделить кластеры сложной формы и открывает новые возможности интерпретации результатов классификации. Более того, следует отметить, что в отличие от вероятностного пространства структура нечетких множеств представлена не булевой решеткой, а векторной, так что с точки зрения формальной логики для нечетких классов не выполняются такие законы, как закон противоречия и закон исключенного третьего.

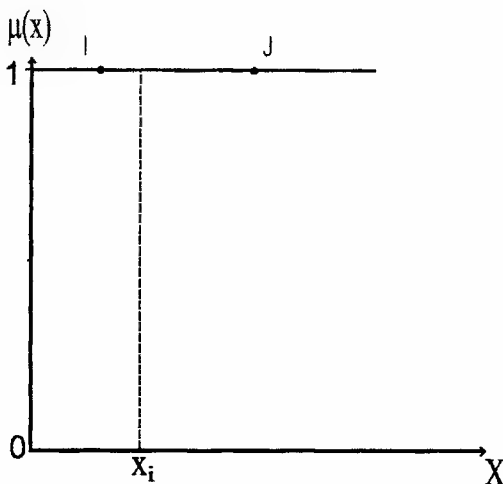


Рис. 1.5. Принадлежность элемента двум четким кластерам

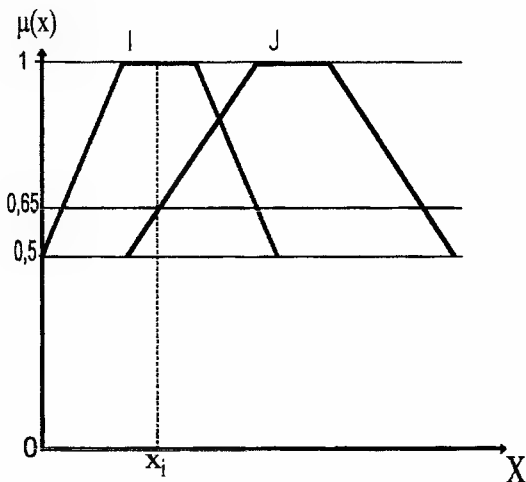


Рис. 1.6. Принадлежность элемента двум нечетким кластерам

Эта особенность теории нечетких множеств также открывает новые возможности в задачах классификации в отличие от традиционных

статистических методов и позволяет использовать более гибкие методы для представления начальной структуры данных, подлежащих классификации.

Возвращаясь к рассмотрению гносеологических аспектов вопроса о применимости теории нечетких множеств к задачам кластер-анализа, необходимо указать на следующее обстоятельство. Для того чтобы современные автоматические системы могли эффективно заменить человека в управлении, к примеру, производством, и результаты работы подобных систем были бы интерпретируемы естественным образом, необходимо, чтобы математическое обеспечение соответствующих автоматических или автоматизированных систем было основано на принципах, используемых человеком в подобных процессах. Однако данный тезис не следует понимать буквально, тем более, когда речь идет о задачах классификации. К примеру, как показывают результаты экспериментов, человек в многомерных пространствах признаков может различать только такие множества объектов, которые отличаются друг от друга линейными решающими функциями. Если способность человека к идентификации зрительных или речевых сигналов формировалась на протяжении длительного процесса эволюции, то его способность к анализу и классификации абстрактной информации формировалась на протяжении последних нескольких тысячелетий, поэтому человек блестяще классифицирует звуки и символы и гораздо хуже — абстрактные объекты. Алгоритмы же автоматической классификации, работая в подобного рода пространствах, должны классифицировать именно абстрактные объекты, то есть решать весьма трудную для человека задачу. Поэтому при разработке алгоритмов кластер-анализа нет необходимости выявлять класс решающих правил, используемых человеком при решении соответствующих задач. Достаточным в таких условиях оказывается воспроизведение свойств человеческого способа мышления и решения задач, а одним из главных таких свойств, как показывают философско-методологические исследования [162], [164], [173], является нечеткость.

Теория нечетких множеств, вводя вместо строгой принадлежности объекта множеству понятие взвешенной принадлежности, позволяет более адекватно представить точки, находящиеся вне типичной части каждого кластера. Далее, как отмечал американский исследователь Э. Г. Распини, «теория нечетких множеств обеспечивает желаемый перенос выборов таксономических решений из дискретного мет-

рического пространства в непрерывное пространство, в котором понятие «похожая классификация» можно определить более содержательно» [39, с.116]. Различные аспекты нечеткой классификации детально исследуются также в работах [48], [111], [139], [179].

Таким образом, краткое логико-гносеологическое рассмотрение проблемы применимости теории нечетких множеств к задачам кластер-анализа показывает, что эта теория оказывается гибким инструментом для представления данных, когда исходная структура оказывается нечеткой, и более адекватным аппаратом для представления непрерывных кластеров, чем теория вероятностей. Кроме того, теория нечетких множеств позволяет воспроизводить человеческий подход к проблеме классификации и открывает богатые возможности для естественной интерпретации результатов процесса классификации [15]. Вместе с тем, несмотря на специфику теории нечетких множеств, использование аппарата нечетких множеств при решении задач распознавания образов с самообучением не приводит к принципиальному изменению общей постановки задачи классификации объектов в условиях отсутствия обучающих выборок, так что имеет смысл рассуждать не о задаче нечеткой автоматической классификации, что подразумевает выделение задач классификации без обучения при размытости структуры классифицируемых объектов и, соответственно, размытости искомой классификации в самостоятельную область научных исследований, а только о *нечеткой модификации задачи автоматической классификации* и, соответственно, о *нечетких методах решения задачи классификации объектов в условиях отсутствия обучающих выборок*, которые и составляют предмет исследования. Подобное различие терминологии позволяет избежать путаницы при определении характера конкретной задачи и, следовательно, способствует выбору наиболее приемлемого метода классификации для решения задачи.

Итак, на содержательном уровне нечеткая модификация задачи автоматической классификации может быть сформулирована следующим образом: *представить исходное множество объектов $X = \{x_1, \dots, x_n\}$, структура которого характеризуется размытостью и информация о котором задана в виде матрицы «объект-свойство» или матрицы «объект-объект», заранее известным либо нет числом однородных, в некотором смысле нечетких классов адекватным образом. Адекватность представления исследуемой совокупности объектов нечеткими классами определяется, в первую очередь, целями*

исследования, видом искомой структуры классификации и содержательными рассмотрениями проблемы классификации в каждом конкретном случае.

Идея и основные принципы применения аппарата теории нечетких множеств к решению задач классификации были выдвинуты в фундаментальной работе Р. Беллмана, Р. Кэлабы и Л. А. Заде [49], а математическая постановка нечеткой модификации задачи кластерного анализа впервые была рассмотрена Э. Г. Распини в работе [149]. Первым систематическим исследованием, посвященным проблеме применения средств нечеткой математики к решению задач распознавания образов, была диссертация Дж. Беждека [51], после чего были предложены самые разнообразные нечеткие методы распознавания образов, в том числе нечеткие методы автоматической классификации. Однако перед рассмотрением схем алгоритмов нечеткого подхода к решению задачи кластер-анализа, нелишним будет обратиться к основным понятиям теории нечетких множеств.

Резюме

Методы выделения однородных групп объектов, иначе именуемые классификацией, условно объединяются в вероятностный, вариативный и структурный, иначе именуемый геометрическим, подходы. Исходные данные в задачах классификации могут быть представлены в виде матрицы $X_{\text{мол}} = [x'_i]$, именуемой матрицей «объект-признак», где x'_i представляет собой значение i -го признака на j -м объекте исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, а также в виде матриц попарных расстояний $d_{\text{мол}} = [d_{ij}]$ или мер близости $r_{\text{мол}} = [r_{ij}]$, имеющих общее обозначение $\rho_{\text{мол}} = [\rho_{ij}]$, где величина ρ_{ij} характеризует отдаленность или близость объектов x_i и x_j исходного множества X , и именуемых матрицами «объект-объект». Таким образом, при интерпретации объектов как точек в соответствующем пространстве признаков задача классификации заключается в разделении совокупности точек $\{X_i\}, i = \overline{1, n}$ в пространстве $I^m(X)$ на однородные классы таким образом, чтобы точки, принадлежащие одному классу, находились бы относительно близко друг от друга, а сами классы различались бы между собой. Полученные в результате разбиения группы именуются кластерами, классами, образами или таксонами, а методы их обнаружения соответственно называются кластерным анализом,

автоматической классификацией, распознаванием образов с самообучением или численной таксономией. Методы автоматической классификации условно объединяются в эвристическое, иерархическое, оптимизационное и аппроксимационное направления, причем иерархические методы, в свою очередь, условно делятся на агломеративные и дивизимные методы. Любой задаче классификации зачастую оказывается свойственна нечеткость, значительно затрудняющая или вообще делающая невозможным получение решения. Понятие нечеткости является общенаучным и может быть определено как внешнее выражение качества внутренней основы явлений, специфика которого заключается в непрерывности перехода от отсутствия проявления к полному выявлению качества предметов, свойств и отношений реального мира, что находит свое отражение в познавательной и мыслительной деятельности индивида. При анализе форм нечеткости, свойственной процессу анализа данных, более уместным представляется использование понятия неопределенности, под которой следует понимать нечеткость соответствия математической постановки задачи анализа данных предметно-содержательной установке на цели исследования и исходным данным. Основными видами неопределенности в задачах автоматической классификации являются неопределенность целей исследования, включающая такие формы, как неясность и неточность, неопределенность характера результатов исследования, формами проявления которой являются неопределенность формы и взаимного расположения кластеров, неоднозначность числа кластеров и размытость природы искомой классификации, а также неопределенность как характеристика исходных данных, имеющая такие формы, как, опять же, неоднозначность числа кластеров, неоднозначность числа объектов исследуемой совокупности и неоднозначность, неточность либо неполнота значений в матрице исходных данных. Каждому из рассмотренных видов и форм неопределенности в задачах автоматической классификации соответствует определенная методология ее формализации или устранения. Краткое логико-гносеологическое рассмотрение вопроса о возможности применения теории нечетких множеств к решению задач кластер-анализа показывает, что эта теория оказывается гибким инструментом для представления данных, когда исходная структура оказывается нечеткой, и более адекватным аппаратом для представления непрерывных кластеров, чем теория вероятностей; кроме того, теория нечетких множеств позволяет воспроизводить человеческий подход к проблеме классификации и открывает богатые

возможности для естественной интерпретации результатов процесса классификации. На содержательном уровне нечеткая модификация задачи автоматической классификации может быть сформулирована следующим образом: представить исходное множество объектов, структура которого характеризуется размытостью и информация о котором задана в виде матрицы «объект-свойство» или матрицы «объект-объект», заранее известным либо нет числом однородных, в некотором смысле, нечетких классов, адекватным образом. Адекватность представления исследуемой совокупности объектов нечеткими классами определяется целями исследования, видом искомой структуры классификации и содержательными рассмотрениями проблемы классификации в каждом конкретном случае.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

2.1. Нечеткие множества

2.1.1. Понятие взвешенной принадлежности и основные определения теории нечетких множеств

Пусть X — некоторое множество элементов и A — некоторое множество, являющееся подмножеством множества X , так что $A \subseteq X$. Тогда то обстоятельство, что некоторый элемент x множества X является также элементом множества A , выражается с помощью записи $x \in A$. Иными словами, элемент x принадлежит множеству A . В противном случае, если элемент x не принадлежит множеству A , используется выражение $x \notin A$. Понятие принадлежности некоторого элемента x множеству A можно также выразить с помощью характеристической функции $\mu_A(x) = \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases}$, значения которой указывают,

является или не является x элементом множества A . Таким образом, множество A задается простым перечислением элементов x таких, что $\mu_A(x) = 1$, поскольку элементы, для которых $\mu_A(x) = 0$, не принадлежат множеству A .

Вместе с тем, зачастую оказывается существенным вопрос не только о принадлежности или непринадлежности некоторого элемента x множеству A , а о том, до какой степени x является элементом множества A . Сама постановка такого вопроса требует обобщения понятия принадлежности элемента множеству, и, как следствие, речь может идти уже о взвешенной принадлежности элемента x множеству A : элемент x может принадлежать множеству A в той или иной степени. Соответствующим образом обобщается понятие характеристической функции $\mu_A(x)$.

Нечеткое множество $A = \{(x, \mu_A(x))\}$ определяется [192] как совокупность упорядоченных пар, составленных из элементов x универсума X , и соответствующих степеней принадлежности функции $\mu_A(x)$, принимающих значения в интервале $[0,1]$. Поскольку функция принадлежности полностью описывает нечеткое множество, то оно может задаваться непосредственно в виде функции принадлежности

$\mu_A : X \rightarrow [0,1]$. Таким образом, нечеткое множество вводится путем расширения двухэлементного множества принадлежностей $\{0,1\}$ до интервала $[0,1]$. Значения функции принадлежности не стоит отождествлять с вероятностью принадлежности элемента x нечеткому множеству A , поскольку степень принадлежности элемента универсума некоторому нечеткому множеству, определенному на этом универсуме, не имеет, как правило, вероятностной природы. Как правило, при записи некоторого нечеткого множества A элементы x , для которых $\mu_A(x)=0$, как и в четком случае, опускают. В дальнейшем определение основных понятий теории нечетких множеств будет основано на изложении [34].

Носителем $Supp(A)$ нечеткого множества A называется множество точек, таких, что $Supp(A) = \{x \in X \mid \mu_A(x) > 0\}$. *Ядром* $Core(A)$ нечеткого множества A называется множество точек, таких, что $Core(A) = \{x \in X \mid \mu_A(x) = 1\}$. Точки $x \in X$, в которых $\mu_A(x) = 0.5$, именуются *точками перехода*.

Множеством уровня α или α -срезом нечеткого множества A , представленным на рис. 2.1, называется обычное подмножество универсума X , которое определяется выражением

$$A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}, \quad (2.1)$$

где $\alpha \in [0,1]$, а *множество строгого уровня α или строгий α -срез* нечеткого множества A , в свою очередь, определяется выражением $A_\alpha^s = \{x \in X \mid \mu_A(x) > \alpha\}$.

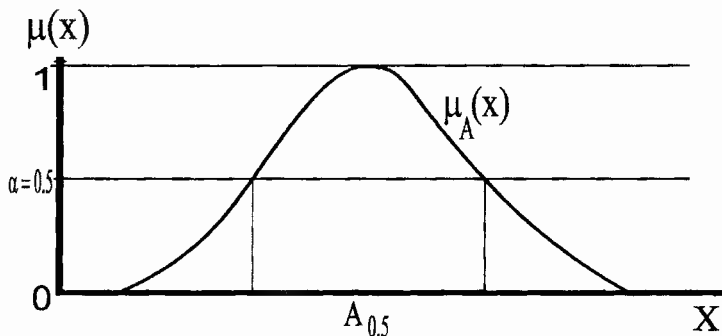


Рис. 2.1. Нечеткое множество и его α -срез

Иными словами, A_α является образом интервала $[\alpha, 1]$ при обратном отображении μ^{-1} , что выражается соотношением

$$A_\alpha = \mu^{-1}([\alpha, 1]). \quad (2.2)$$

Таким образом, носитель $Supp(A)$ нечеткого множества A можно определить как множество строгого уровня 0, а ядро $Core(A)$ нечеткого множества A , соответственно, как множество уровня 1. В соответствии с теоремой декомпозиции [27, с.44], любое нечеткое множество можно разложить по его α -срезам следующим образом:

$$A = \bigcup_{\alpha \in [0,1]} (\alpha \cdot \mu_{A_\alpha}(x)), \text{ где } \mu_{A_\alpha}(x) = \begin{cases} 1, \text{ если } \mu_A(x) \geq \alpha \\ 0, \text{ если } \mu_A(x) < \alpha \end{cases}. \quad (2.3)$$

Нечеткое множество уровня α определяется [147] как множество упорядоченных пар

$$A_{(\alpha)} = \{x \in A_\alpha, \mu_{A_{(\alpha)}}(x) = \mu_A(x)\}, \alpha \in [0,1], \quad (2.4)$$

где $A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}$.

Обычное множество, ближайшее к нечеткому, определяется следующим образом:

$$\mu_A(x) = \begin{cases} 0, \mu_A(x) < 0.5 \\ 1, \mu_A(x) > 0.5 \\ 0 \text{ или } 1, \mu_A(x) = 0.5 \end{cases}. \quad (2.5)$$

Высотой нечеткого множества A называется величина $\sup_{x \in X} \mu_A(x)$. Нечеткое множество A называется *нормальным*, если $\sup_{x \in X} \mu_A(x) = 1$; в случае, когда $\sup_{x \in X} \mu_A(x) < 1$, нечеткое множество A называется *субнормальным*. Субнормальное нечеткое множество A можно *нормализовать* в соответствии с формулой

$$\tilde{\mu}_A(x) = \frac{\mu_A(x)}{\sup_{x \in X} \mu_A(x)}. \quad (2.6)$$

Если X — линейное векторное пространство, то нечеткое множество A будет *выпуклым* тогда и только тогда, когда его функция принадлежности выпукла, то есть для каждой пары точек $x_i, x_j \in X$ функция принадлежности нечеткого множества A удовлетворяет неравенству

$$\mu_A(\lambda x_i + (1-\lambda)x_j) \geq \min(\mu_A(x_i), \mu_A(x_j)) \quad (2.7)$$

для всех $\lambda: 0 \leq \lambda \leq 1$.

Понятие нечеткого множества может обобщаться различным образом [27], [34, с.19-29]. Одним из обобщений понятия нечеткого множества является понятие нечеткого множества типа p . Приведенное выше определение нечеткого множества, в соответствии с которым функция принадлежности нечеткого множества A на универсуме X определяется выражением $\mu_A : X \rightarrow [0,1]$, задает нечеткое множество типа 1. Нечеткое множество типа $p, p=1,2,\dots$ есть нечеткое множество, значениями функции принадлежности которого являются нечеткие множества типа $(p-1)$.

Для записи нечетких множеств используются различные формы. К примеру, Л. А. Заде использует следующие обозначения [24]. Если универсум $X = \{x_1, \dots, x_n\}$ записывать в виде

$$X = x_1 + \dots + x_n, \quad (2.8)$$

или

$$X = \sum_{i=1}^n x_i, \quad (2.9)$$

где символ $+$ обозначает объединение, а не арифметическое суммирование, то нечеткое множество A универсума X будет записываться в виде

$$A = \mu_1 / x_1 + \dots + \mu_n / x_n \quad (2.10)$$

или, соответственно, в виде

$$A = \sum_{i=1}^n \mu_i / x_i, \quad (2.11)$$

где $\mu_i, i=1, \dots, n$ представляют собой степени принадлежности элементов x_i нечеткому множеству A , а символ $/$ используется для различения компонент μ_i и x_i во избежание неопределенности, так что записи (2.10), (2.11) можно интерпретировать как представление нечеткого множества A в виде объединения составляющих его одноточечных нечетких множеств μ_i / x_i .

Если носитель $Supp(A)$ нечеткого множества A имеет мощность континуума, то используется запись

$$A = \int_x \mu_A(x) / x, \quad (2.12)$$

где $\mu_A(x)$ — функция принадлежности нечеткого множества A , а символ $\int_x$ обозначает объединение нечетких синглетонов $\mu_A(x) / x$ по рассматриваемому универсуму X .

На рис. 2.2 представлены некоторые возможные виды функции принадлежности нечетких множеств с континуальным носителем на универсуме X : унимодальная (U), мультимодальная (M), субнормальная (S), амодальная (A).

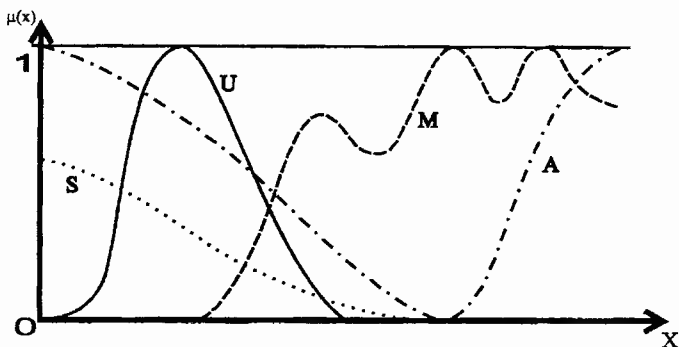


Рис. 2.2. Виды функции принадлежности нечетких множеств

Достаточно полный обзор функций принадлежности содержится в [27].

2.1.2. Основные операции над нечеткими множествами

Пусть A и B — два нечетких множества, определенные на универсуме X . Для определения основных операций над нечеткими множествами целесообразно использование традиционной формы записи, преимущественно используемой в литературе по нечетким множествам.

Включение нечеткого множества A в нечеткое множество B , которое представлено на рис. 2.3, определяется условием $\forall x \in X : \mu_A(x) \leq \mu_B(x)$.

Равенство нечеткого множества A нечеткому множеству B , которое представлено на рис. 2.4, определяется условием $\forall x \in X : \mu_A(x) = \mu_B(x)$.

Дополнение нечеткого множества A , представленное на рис. 2.5, определяется условием $\forall x \in X : \mu_{\bar{A}}(x) = 1 - \mu_A(x)$.

В теории нечетких множеств для обозначения операции взятия минимума используется символ $\wedge$, а для обозначения операции взятия максимума — символ $\vee$.

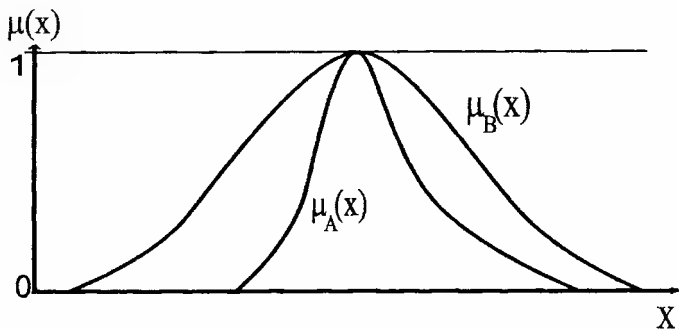


Рис. 2.3. Включение одного нечеткого множества в другое

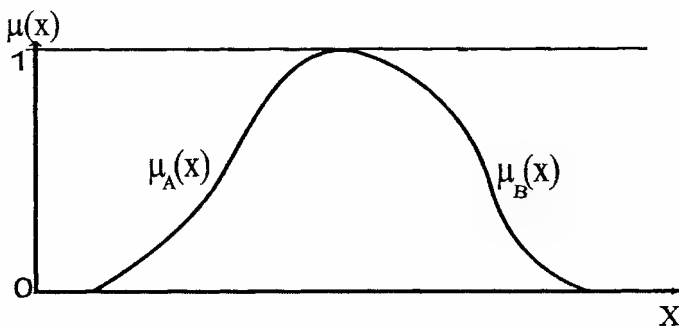


Рис. 2.4. Равенство нечетких множеств

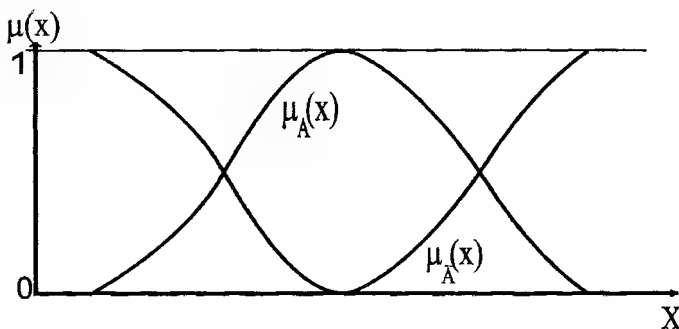


Рис. 2.5. Дополнение нечеткого множества

Пересечение нечетких множеств A и B , представленное на рис. 2.6, определяется как наибольшее нечеткое подмножество, содержащееся одновременно и в нечетком множестве A , и в нечетком множестве B : $\forall x \in X: \mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) = \mu_A(x) \wedge \mu_B(x)$.

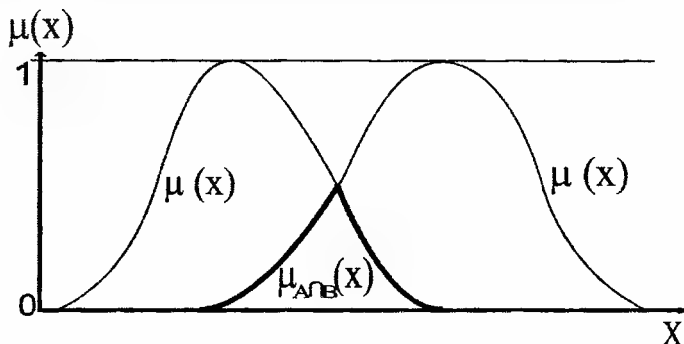


Рис. 2.6. Пересечение нечетких множеств

Объединение нечетких множеств A и B , представленное на рис. 2.7, определяется как наименьшее нечеткое подмножество, содержащееся одновременно и в нечетком множестве A , и в нечетком множестве B : $\forall x \in X: \mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) = \mu_A(x) \vee \mu_B(x)$.

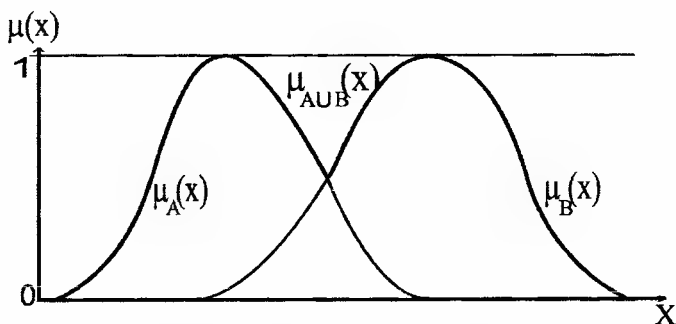


Рис. 2.7. Объединение нечетких множеств

В работе [27] также рассматривается обобщение понятия расстояния на случай нечетких множеств. В частности, *относительное*

обобщенное расстояние Хемминга между нечеткими множествами A и B определяется по формуле

$$d_H(A, B) = \frac{1}{n} \sum_{i=1}^n |\mu_A(x_i) - \mu_B(x_i)|, \quad (2.13)$$

где $n = \text{card}(X)$, $x_i \in A$, $x_i \in B$, $A, B \subseteq X$, $\mu_A(x_i), \mu_B(x_i) \in [0, 1]$, $i = 1, \dots, n$, так что очевидно, что $0 \leq d_H(A, B) \leq 1$. Относительное евклидово расстояние между нечеткими множествами A и B определяется по формуле

$$d_E(A, B) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_A(x_i) - \mu_B(x_i))^2}, \quad (2.14)$$

где $n = \text{card}(X)$, $x_i \in A$, $x_i \in B$, $A, B \subseteq X$, $\mu_A(x_i), \mu_B(x_i) \in [0, 1]$, $i = 1, \dots, n$, так что $0 \leq d_E(A, B) \leq 1$.

Более полный обзор операций над нечеткими множествами содержится в [27], [34].

2.2. Нечеткие отношения

2.2.1. Определение нечеткого отношения; свойства и классификация нечетких отношений

Бинарным нечетким отношением R между множествами X и Y будет называться функция $\mu_R : X \times Y \rightarrow [0, 1]$ где μ_R — функция принадлежности отношения R ; X и Y — базовые множества; $[0, 1]$ — отрезок вещественной прямой от 0 до 1. В случае, когда множества X и Y в силу связи между объектами совпадают, нечеткое отношение $\mu_R : X \times X \rightarrow [0, 1]$, где X — базовое множество, или, иначе говоря, универсум, называется нечетким отношением на множестве X . Другими словами, бинарное нечеткое отношение является нечетким множеством на декартовом произведении базовых множеств. В дальнейшем, за некоторыми исключениями, которые будут оговариваться особо, рассматриваются нечеткие отношения на множестве X . Поскольку нечеткое отношение может определяться как нечеткое множество на декартовом произведении базовых множеств, так что нечеткое отношение R на множестве X может задаваться непосредственно в виде функции принадлежности $R = \{\mu_R(x_i, x_j) | x_i, x_j \in X\}$, для нечетких отношений сохраняются многие результаты, полученные для нечетких множеств. В частности, понятие нечеткого отношения

также может обобщаться, и одним из обобщений понятия нечеткого отношения является понятие нечеткого отношения типа p .

В зависимости от свойств, которыми они обладают, выделяют различные типы нечетких отношений. Среди всех рассматриваемых в специальной литературе свойств нечетких отношений, полагая, что S — некоторое произвольное бинарное нечеткое отношение на множестве X , следует особо отметить следующие свойства:

рефлексивность:

$$\mu_s(x_i, x_i) = 1, \forall x_i \in X, \quad (2.15)$$

слабая рефлексивность:

$$\mu_s(x_i, x_j) \leq \mu_s(x_j, x_j), \forall x_i, x_j \in X, \quad (2.16)$$

сильная рефлексивность:

$$\mu_s(x_i, x_i) = 1, \mu_s(x_i, x_j) < 1, \forall x_i, x_j \in X, x_i \neq x_j, \quad (2.17)$$

антирефлексивность:

$$\mu_s(x_i, x_i) = 0, \forall x_i \in X, \quad (2.18)$$

слабая антирефлексивность:

$$\mu_s(x_i, x_j) \geq \mu_s(x_j, x_j), \forall x_i, x_j \in X, \quad (2.19)$$

сильная антирефлексивность:

$$\mu_s(x_i, x_i) = 0, \mu_s(x_i, x_j) > 0, \forall x_i, x_j \in X, i \neq j, \quad (2.20)$$

симметричность:

$$\mu_s(x_i, x_j) = \mu_s(x_j, x_i), \forall x_i, x_j \in X, \quad (2.21)$$

антисимметричность:

$$\mu_s(x_i, x_j) \wedge \mu_s(x_j, x_i) = 0, \forall x_i, x_j \in X, i \neq j, \quad (2.22)$$

асимметричность:

$$\mu_s(x_i, x_j) \wedge \mu_s(x_j, x_i) = 0, \forall x_i, x_j \in X, \quad (2.23)$$

полнота сильная:

$$\mu_s(x_i, x_j) \vee \mu_s(x_j, x_i) = 1, \forall x_i, x_j \in X, \quad (2.24)$$

полнота слабая:

$$\mu_s(x_i, x_j) \vee \mu_s(x_j, x_i) > 0, \forall x_i, x_j \in X, \quad (2.25)$$

(max-min)-транзитивность:

$$\mu_s(x_i, x_k) \geq \bigvee_{x_j \in X} (\mu_s(x_i, x_j) \wedge \mu_s(x_j, x_k)), \forall x_i, x_j, x_k \in X, \quad (2.26)$$

(min-max)-транзитивность:

$$\mu_s(x_i, x_k) \leq \bigwedge_{x_j \in X} (\mu_s(x_i, x_j) \vee \mu_s(x_j, x_k)), \forall x_i, x_j, x_k \in X. \quad (2.27)$$

Очевидно, к примеру, что если нечеткое отношение обладает свойством (2.17), то оно обладает и свойством (2.15), так же, как на-

личие у нечеткого отношения свойства (2.20) подразумевает и наличие у него свойства (2.18). Наиболее полно свойства нечетких отношений рассмотрены в работах [27], [28], [34]. Классификация нечетких отношений может быть представлена, как и в работе [27], в виде таблицы 2.1.

Таблица 2.1

Классификация основных нечетких отношений

Тип нечеткого отношения	Свойства нечеткого отношения					
	2.15	2.18	2.26	2.27	2.21	2.22
Подобие	+		+		+	
Различие		+		+	+	
Толерантность	+				+	
Несходство		+			+	
Предпорядок	+		+			
Слабый порядок	+					+
Нестрогий порядок	+		+			+
Строгий порядок		+	+			+

Более подробная классификация нечетких отношений приводится в работе [34]. Следует отметить, что различные авторы, рассматривая разные типы нечетких отношений, используют различные названия и обозначения. Помимо вышеприведенных нечетких отношений, особо следует отметить нечеткое отношение моделирования [1], [2], определяющее соответствие между внешней средой и внутренней моделью внешнего мира, а также отношение похожести, предложенное Э. Г. Распини [39] и применяемое при рассмотрении теоретических аспектов нечеткого подхода в кластер-анализе.

В рамках нечеткого подхода к решению задач автоматической классификации традиционно применяются нечеткие отношения эквивалентности, именуемые также отношениями подобия и традиционно обозначаемые символом S , а также нечеткие отношения схождения, называемые еще нечеткими толерантностями, нечеткими отношениями безразличия или нечеткими отношениями неразличимости и обозначаемые символом T . Кроме того, в задачах автоматической классификации применяются также нечеткие отношения различия, являющиеся дополнениями нечетких отношений эквивалентности и обозначаемые символом D , а также нечеткие отношения несхождения, являющиеся дополнениями нечетких толерантностей и, в свою очередь, обозначаемые символом I . Подробное рассмотрение вопросов, относящихся

к применению нечетких отношений в задачах классификации, содержится в работе [190].

В работах [174], [175], [16] были предложены нечеткие отношения сильной, слабой и строгой слабой толерантностей, а в работе [177] были рассмотрены их дополнения, именуемые соответственно нечеткими отношениями сильного, слабого и строгого слабого несходства. Строгая слабая толерантность отличается от слабой толерантности выполнением строгого неравенства в условии (2.16):

$$\mu_s(x_i, x_j) < \mu_s(x_j, x_j), \forall x_i, x_j \in X, \quad (2.28)$$

а строгое слабое несходство отличается от слабого несходства, соответственно, выполнением строгого неравенства в условии (2.19):

$$\mu_s(x_i, x_j) > \mu_s(x_j, x_j), \forall x_i, x_j \in X. \quad (2.29)$$

Основные свойства, по которым выделяют различные нечеткие толерантности, а также нечеткие отношения несходства, представлены таблицей 2.2.

Таблица 2.2

Нечеткие толерантности и нечеткие отношения несходства, применяемые в кластер-анализе

Тип нечеткого отношения и его обозначение	Свойства нечетких отношений								
	2.15	2.16	2.17	2.18	2.19	2.20	2.21	2.28	2.29
Обычная толерантность T_2	+						+		
Сильная толерантность T_3			+				+		
Слабая толерантность T_1		+					+		
Строгая слабая толерантность T_0							+	+	
Обычное несходство I_2				+			+		
Сильное несходство I_3						+	+		
Слабое несходство I_1					+		+		
Строгое слабое несходство I_0							+		+

Нечеткому отношению R на множестве X можно поставить в соответствие взвешенный граф, где каждая пара вершин (x_i, x_j) из X соединяется дугой с весом $\mu_R(x_i, x_j)$. Очевидно, что для симметричных отношений такой граф будет неориентированным. В качестве иллюстрации целесообразно привести пример, представленный в работе [163]. Пусть $X = \{x_1, x_2, x_3, x_4, x_5\}$ — множество элементов, на котором задана нечеткая толерантность T , представленная таблицей 2.3. Соответствующий ей граф изображен на рис. 2.8.

Таблица 2.3

Нечеткая толерантность T
на множестве X

T	x_1	x_2	x_3	x_4	x_5
x_1	1				
x_2	0.8	1			
x_3	0	0.4	1		
x_4	0.1	0	0	1	
x_5	0.2	0.9	0	0.5	1

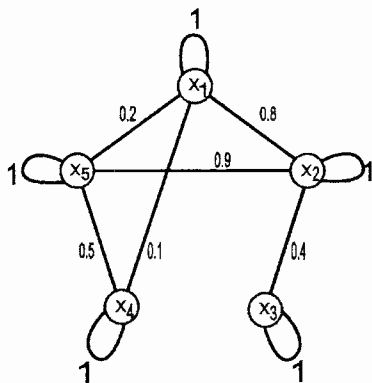


Рис. 2.8. Граф сходства нечеткой толерантности T

Необходимо отметить, что представленная в таблице 2.3 и на рис. 2.8 нечеткая толерантность T является нечеткой сильной толерантностью T_3 .

В завершение рассмотрения основных понятий и определений нечетких отношений следует указать, что нечеткие отношения слабой и строгой слабой толерантностей могут применяться для представления сходства динамических объектов [176]. В задачах кластерного анализа преимущественно используются нечеткие сильные и обычные толерантности, для которых выполняется соотношение $T_3 \subseteq T_2$, и нечеткие отношения сильного и обычного несходства, для которых, соответственно, выполняется условие $I_2 \subseteq I_3$. В дальнейшем нечеткие сильные и обычные толерантности будут обозначаться символом T

без индекса, а нечеткие отношения сильного и обычного несходства — соответственно, символом I без индекса.

Свойства и содержательная интерпретация нечетких эквивалентностей и нечетких толерантностей, а также их дополнений подробно рассматриваются в работах [12], [27], [28], [34], [67], [68], [193].

2.2.2. Основные операции над нечеткими отношениями

Пусть R и S — два произвольных бинарных нечетких отношения на множестве X .

Включение нечеткого отношения R в нечеткое отношение S определяется условием

$$R \subseteq S \Leftrightarrow \mu_R(x_i, x_j) \leq \mu_S(x_i, x_j), \forall x_i, x_j \in X. \quad (2.30)$$

Дополнение $\bar{R}$ нечеткого отношения R определяется условием

$$\mu_{\bar{R}}(x_i, x_j) = 1 - \mu_R(x_i, x_j), \forall x_i, x_j \in X. \quad (2.31)$$

Пересечение нечетких отношений R и S определяется условием

$$\mu_{R \cap S}(x_i, x_j) = \mu_R(x_i, x_j) \wedge \mu_S(x_i, x_j), \forall x_i, x_j \in X. \quad (2.32)$$

Объединение нечетких отношений R и S определяется условием

$$\mu_{R \cup S}(x_i, x_j) = \mu_R(x_i, x_j) \vee \mu_S(x_i, x_j), \forall x_i, x_j \in X. \quad (2.33)$$

Пусть X, Y, Z — универсальные множества, так что нечеткое отношение R определено на $X \times Y$, а нечеткое отношение S — соответственно на $Y \times Z$, тогда *(max-min)-композицией* нечетких отношений R и S , $\mu_R : X \times Y \rightarrow [0, 1]$, $\mu_S : Y \times Z \rightarrow [0, 1]$ называется отношение $R \circ S$, определяемое условием

$$\mu_{R \circ S}(x_i, x_k) = \bigvee_{x_j \in Y} (\mu_R(x_i, x_j) \wedge \mu_S(x_j, x_k)), \forall x_i \in X, x_k \in Z. \quad (2.34)$$

Свойство транзитивности нечетких отношений (2.26) может быть выражено, таким образом, в виде условия

$$R \supseteq R \circ R. \quad (2.35)$$

Если в условии (2.34) операцию взятия минимума $\wedge$ заменить на операцию взятия среднего арифметического, получится следующее определение операции композиции:

$$\mu_{R \circ S}(x_i, x_k) = \bigvee_{x_j \in Y} (0.5(\mu_R(x_i, x_j) + \mu_S(x_j, x_k))), \forall x_i \in X, x_k \in Z, \quad (2.36)$$

именуемое *(max-sum)-композицией*. Возможны также и другие определения операции композиции нечетких отношений; например, в работе [9] рассмотрена операция *(max-•)-композиции*:

$$\mu_{R \circ S}(x_i, x_k) = \bigvee_{x_j \in Y} (\mu_R(x_i, x_j) \cdot \mu_S(x_j, x_k)), \forall x_i \in X, x_k \in Z, \quad (2.37)$$

полученная заменой в условии (2.34) операции взятия минимума $\wedge$ на операцию умножения.

Как и при рассмотрении нечетких множеств, при рассмотрении нечетких отношений вводится понятие α -среза, так что α -срезом R_α некоторого бинарного нечеткого отношения R называется обычное отношение на универсуме X , которое определяется выражением

$$R_\alpha = \{(x_i, x_j) \in X \times X \mid \mu_R(x_i, x_j) \geq \alpha\}, \quad (2.38)$$

где $\alpha \in [0,1]$, что, в свою очередь, в соответствии с теоремой декомпозиции [27, с.75], позволяет представить нечеткое отношение R в виде иерархии обычных отношений:

$$R = \bigvee_{\alpha \in (0,1]} (\alpha \cdot \mu_{R_\alpha}(x_i, x_j)), \text{ где } \mu_{R_\alpha}(x_i, x_j) = \begin{cases} 1, \text{ если } \mu_R(x_i, x_j) \geq \alpha \\ 0, \text{ если } \mu_R(x_i, x_j) < \alpha \end{cases}. \quad (2.39)$$

Естественно, что смысл термина «декомпозиция» в этом случае отличается от смысла данного термина, когда речь идет о композиции нечетких отношений. В работе [9] подробно рассматриваются особенности операции декомпозиции нечетких отношений эквивалентности, которая находит обширное применение в задачах кластерного анализа. Поскольку большое значение в приложениях теории нечетких отношений играют транзитивные нечеткие отношения, необходимо рассмотреть операцию преобразования исходного нетранзитивного отношения в транзитивное. Данная операция, именуемая транзитивным замыканием, впервые была рассмотрена в работах [163], [193].

(*max-min*)-транзитивным замыканием бинарного нечеткого отношения R на множестве X , где $\text{card}(X) = n$, называется бинарное нечеткое отношение $\tilde{R}$ на множестве X , определяемое следующим образом:

$$\tilde{R} = R^1 \cup R^2 \cup \dots \cup R^n, \quad (2.40)$$

где отношения R^n определяются рекурсивно:

$$R^1 = R, R^n = R^{n-1} \circ R, n = 2, 3, \dots \quad (2.41)$$

Иллюстративным примером данной операции может послужить нечеткое отношение, представленное таблицей 2.4 и являющееся (*max-min*)-транзитивным замыканием нечеткой толерантности T , представленной таблицей 2.3. Соответствующий (*max-min*)-транзитивному замыканию $\tilde{T}$ нечеткой толерантности T граф изображен на рис. 2.9.

Таблица 2.4
(max-min)-транзитивное замыкание
 $\tilde{T}$ нечеткой толерантности T

$\tilde{T}$	x_1	x_2	x_3	x_4	x_5
x_1	1				
x_2	0.8	1			
x_3	0.4	0.4	1		
x_4	0.5	0.5	0.4	1	
x_5	0.8	0.9	0.4	0.5	1

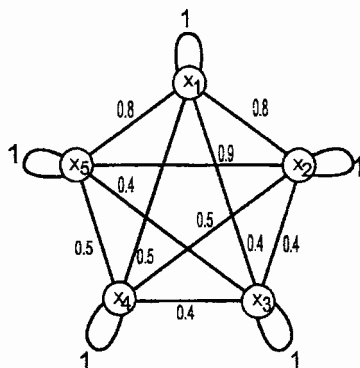


Рис. 2.9. Граф нечеткого отношения $\tilde{T}$

В работе [34] формулируется теорема, в соответствии с которой транзитивное замыкание $\tilde{R}$ любого бинарного нечеткого отношения R транзитивно и является наименьшим транзитивным отношением, включающим R , то есть $R \subseteq \tilde{R}$, и для любого транзитивного отношения S , такого, что $R \subseteq S$, следует $\tilde{R} \subseteq S$.

Следует отметить [34, с.45], что при транзитивном замыкании некоторого нечеткого отношения из свойств (2.15) — (2.27) сохраняются только свойства (2.15), (2.21), (2.24), (2.25), (2.26), (2.27).

(min-max)-транзитивным замыканием [27, с.137-138] бинарного нечеткого отношения R на множестве X , где $\text{card}(X) = n$, называется бинарное нечеткое отношение $\hat{R}$ на множестве X , определяемое следующим образом:

$$\hat{R} = R \cap (R * R) \cap \dots \cap (R * \dots_n R), \quad (2.42)$$

где отношения $(R * \dots_n R)$ определяются рекурсивно:

$$\mu_{R * R}(x_i, x_k) = \bigwedge_{x_j \in X} (\mu_R(x_i, x_j) \vee \mu_R(x_j, x_k)), \forall x_i, x_j, x_k \in X. \quad (2.43)$$

Полное рассмотрение операций над нечеткими отношениями содержится в [27], а применение аппарата нечетких графов к решению задач кластерного анализа рассматривается в работе [190].

Резюме

Понятие нечеткого множества вводится путем расширения двухэлементного множества принадлежностей $\{0,1\}$ элементов универсума X его подмножеству A до интервала $[0,1]$, так что нечеткое множество $A = \{(x, \mu_A(x))\}$ определяется как совокупность упорядоченных пар, составленных из элементов x универсума X , и соответствующих степеней принадлежности функции $\mu_A(x)$, принимающих значения в интервале $[0,1]$. Поскольку функция принадлежности полностью описывает нечеткое множество, то оно может задаваться непосредственно в виде функции принадлежности $\mu_A : X \rightarrow [0,1]$. Понятие нечеткого множества может различным образом обобщаться, и одним из обобщений является понятие нечеткого множества типа p . Над нечеткими множествами определяются такие основные операции, как включение, равенство, дополнение, пересечение, объединение и некоторые другие, не имеющие аналогов в случае обычных множеств, а также понятие расстояния между нечеткими множествами. Бинарным нечетким отношением R между множествами X и Y называется функция $\mu_R : X \times Y \rightarrow [0,1]$; в случае, когда множества X и Y совпадают, нечеткое отношение $\mu_R : X \times X \rightarrow [0,1]$ называется бинарным нечетким отношением R на множестве X . Иначе, нечеткое отношение может определяться как нечеткое множество на декартовом произведении базовых множеств, так что для нечетких отношений сохраняются многие результаты, полученные для нечетких множеств. В зависимости от свойств, которыми они обладают, выделяют различные типы нечетких отношений, такие как отношения подобия, различия, толерантности, несходства, предпорядка, слабого порядка, строгого порядка, нестрогого порядка и некоторые другие. Любому нечеткому отношению R на множестве X можно поставить в соответствие взвешенный граф, где каждая пара вершин (x_i, x_j) из X соединяется дугой с весом $\mu_R(x_i, x_j)$. Над нечеткими отношениями также определяются различные операции, как свойственные четким отношениям, так и не имеющие аналога в четком случае.

МЕТОДЫ НЕЧЕТКОГО ПОДХОДА К РЕШЕНИЮ ЗАДАЧ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

3.1. Эвристические методы

3.1.1. Особенности эвристических методов решения нечеткой модификации задачи автоматической классификации

Эвристические алгоритмы, как правило, непосредственно опираются на постановку задачи выделения в исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, структура которой характеризуется нечеткостью, групп объектов $A^1, \dots, A^c$, являющихся нечеткими множествами с соответствующими функциями принадлежности $\mu_{A^l}(x_i), l = 1, \dots, c, i = 1, \dots, n$, или, для краткости, $\mu_{li}, l = 1, \dots, c, i = 1, \dots, n$. В основе такой постановки, как правило, лежат содержательные представления исследователя о нечеткости структуры исследуемой совокупности и условиях объединения объектов в классы. Число нечетких эвристических кластер-процедур немногочисленно, и, пожалуй, единственной особенностью, общей для всех этих алгоритмов, является то обстоятельство, что в качестве входных данных в них используется матрица вида «объект-объект»; исключением является только алгоритм, предложенный И. Гитманом и М. Д. Левиным [90], где исходные данные задаются в виде матрицы «объект-признак» (1.1). Таким образом, основные определения и особенности каждого алгоритма целесообразно рассматривать по ходу его описания в каждом отдельном случае, а перед этим ограничиться краткой характеристикой каждого из представленных ниже алгоритмов.

Наиболее интересной особенностью алгоритма классификации, предложенного И. Гитманом и М. Д. Левиным [90], является то обстоятельство, что группировка объектов производится с использованием как значений функции принадлежности нечеткого множества объектов, подлежащих классификации, так и некоторой заданной метрики, что позволяет выделять классы объектов достаточно сложной формы при отсутствии априорной информации о числе классов. Вместе с тем, необходимо указать, что в данной процедуре понятие нечеткого множества, вообще говоря, используется лишь как вспомога-

тельное, так что задача автоматической классификации решается, по сути, в детерминистском смысле, что нашло свое отражение в полном отказе от использования понятия нечеткого множества в более поздней версии алгоритма [91]. Подобные соображения приводятся также в работе И. И. Елисеевой и В. О. Рукавишникова [22, с.53].

Метод классификации, предложенный С. Тамурой, С. Хигути и К. Танакой в работе [163], использует операцию (max-min)-транзитивного замыкания (2.40) нечеткой толерантности T , описывающей исходные данные в виде матрицы близости $r_{\text{пол}} = [\mu_T(x_i, x_j)], i = 1, \dots, n, j = 1, \dots, n$, а процедура, соответствующая данному методу, оказывается весьма простой с вычислительной точки зрения и строит иерархию четких разбиений по α -уровням транзитивного замыкания $\tilde{T}$. Таким образом, для выбора некоторого единственного разбиения требуется априорная информация о числе классов. Если некоторое разбиение выбрано, то соответствующее значение α будет представлять собой порог сходства элементов, объединяемых в кластеры, что является довольно существенным обстоятельством при содержательной интерпретации полученного результата. Детальное рассмотрение некоторых результатов, представленных в исследовании [163], было произведено также Дж. Данном в работах [82], [86].

А. Кутюрье и Б. Фьолео в работе [71] предложили достаточно простой с вычислительной точки зрения алгоритм, строящий покрытие исследуемой совокупности $X = \{x_1, \dots, x_n\}$ объектов нечеткими кластерами. Нечеткие множества $A^l, l = 1, \dots, c$, определенные на универсуме $X = \{x_1, \dots, x_n\}$, с функциями принадлежности $\mu_{li}, l = 1, \dots, c, i = 1, \dots, n$, образуют покрытие $C = \{A^1, \dots, A^c\}$, если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^c \mu_{li} \geq 1$. Таким образом, для данного метода оказывается характерной возможность пересечения нечетких кластеров, что позволяет углубить анализ результатов нечеткой классификации. Однако следует отметить, что в предложенном в работе [71] методе понятие нечеткого кластера достаточно точно определено, так что не любое нечеткое множество, определенное на множестве объектов $X = \{x_1, \dots, x_n\}$, рассматривается как нечеткий кластер.

В алгоритме классификации, предложенном Л. С. Берштейном и Т. А. Дзюбой [11], исходные данные представляются в виде нечеткого графа, и алгоритм представляет собой процедуру разбиения множеств

ва объектов $X = \{x_1, \dots, x_n\}$, являющихся вершинами нечеткого графа, на два класса путем выделения в графе такой двудольной структуры, при которой введенная в качестве критерия оптимизации разбиения степень двудольности является максимальной. В силу этого обстоятельства данную процедуру можно отнести к алгоритмам эвристического направления лишь с некоторой долей условности, поскольку, как отмечает И. Д. Мандель, «проблематика разрезания графа в смысле некоторых оптимизационных требований является специфическим направлением теории графов» [31, с.75].

Нечеткие кластер-процедуры эвристического направления, помимо решения собственно задач классификации, имеют большое значение на этапе предварительного анализа данных, когда неизвестны число кластеров, их структура и взаимное расположение. К примеру, в работе [22] алгоритм Гитмана-Левина рассматривается как процедура, предвещающая работу нечеткой оптимизационной кластер-процедуры, предложенной Э. Г. Распини [150], с целью определения числа и центров классов, а также построения первоначального разбиения.

3.1.2. Описание алгоритмов

Различные нечеткие эвристические кластер-процедуры в качестве основы используют различные понятия и определения, так что их необходимо рассматривать перед изложением схемы алгоритма в каждом конкретном случае.

3.1.2.1. Алгоритм Гитмана — Левина [90] является одной из первых эвристических процедур, использующих понятие нечеткого множества, и строит разбиение нечеткого множества объектов, подлежащих классификации, на унимодальные нечеткие множества.

Основные понятия и определения

Пусть $X = \{x_1, \dots, x_n\}$ — множество классифицируемых объектов, на котором задано нечеткое множество $B = \{(x_i, \mu_B(x_i))\}, i = 1, \dots, n$; таким образом, B — дискретное нечеткое множество точек, подлежащих классификации. Пусть τ — точка, в которой значение функции принадлежности максимально, то есть $\mu_B(\tau) = \sup_{x_i \in X} \mu_B(x_i)$. В таком случае на множестве B могут быть определены следующие множества:

$$B_g = \{x_i \mid \mu_B(x_i) \geq \mu_B(x_g)\} \quad (3.1)$$

и

$$B_{g(d)} = \{x_i \mid d(\tau, x_i) \leq d(\tau, x_g)\}, \quad (3.2)$$

где $x_i \in B$ и d — некоторая метрика. Необходимо отметить, что при $\mu_B(x_g) = \alpha, \alpha \in [0, 1]$ четкое множество B_g представляет собой множество уровня α нечеткого множества B в смысле условия (2.1).

Нечеткое множество B является *симметричным* в том и только в том случае, если $\forall x_i \in B, B_g = B_{g(d)}$. Очевидно, что если нечеткое множество B является симметричным, то справедливо соотношение $\forall x_i, x_k \in B, d(x_i, \tau) \leq d(x_k, \tau) \Leftrightarrow \mu_B(x_i) \geq \mu_B(x_k)$.

Нечеткое множество B является *унимодальным* в том и только в том случае, если четкое множество B_g является связанным $\forall x_i \in B$.

Дискретные нечеткие множества $\tilde{A}^1, \dots, \tilde{A}^G$ определяют разбиение нечеткого множества B на G нечетких подмножеств. Точка $\hat{\tau}^s \in \tilde{A}^s, s \in [1, G]$, для которой выполняется условие $\mu_{\tilde{A}^l}(\hat{\tau}^l) = \sup_{x_i \in \tilde{A}^l} \mu_{\tilde{A}^l}(x_i), l = 1, \dots, G$, будет именоваться *модальной точкой*

или просто *модой* группы $\tilde{A}^s$. Некоторая мода $\hat{\tau}^s \in \tilde{A}^s, s = 1, \dots, G$ будет именоваться *локальным максимумом* функции $\mu_B(x_i)$ и обозначаться символом τ^s , если для нее выполняется соотношение $\mu_B(\tau^s) = \max_{x_i \in B} \mu_B(x_i)$.

Некоторой точке $x_k \in \tilde{A}^s$ ставится в соответствие точка $x_i \in (B - \tilde{A}^s)$, такая, что $d(x_i, x_k) = \min_{x_j \in (B - \tilde{A}^s)} [d(x_k, x_j)]$, тогда точка x_k будет именоваться *внутренней точкой* подмножества $\tilde{A}^s$ в том и только в том случае, если множество $B_j = \{x_i \mid d(x_j, x_i) < d(x_j, x_k)\}$ включает в себя по меньшей мере одну точку из множества $\tilde{A}^s$. Необходимо указать, что в оригинальной версии алгоритма [90, с.588] используется пороговое значение $\hat{d}, \hat{d} > 0$, так что $B_j = \{x_i \mid d(x_j, x_i) \leq \hat{d}\}$, и с каждой точкой x_j сопоставляется число $u_j = \text{card}(B_j)$; иными словами, u_j — число точек в $\hat{d}$ -окрестности точки x_j [22, с.54]. Очевидно, что если в качестве функции, сопоставляющей точке x_j значение u_j выступает функция принадлежности нечеткого множества B , то в качестве порога может быть выбрано некоторое значение α — порог различия элементов, объединяемых в кластеры, $\alpha \in [0, 1]$. Следует отме-

титель, что результат разбиения на группы $\tilde{A}^1, \dots, \tilde{A}^c$ полностью определяется порогом $\hat{d}$, который должен быть не настолько низким, что $u_j = 1, \forall x_j \in X$, и не настолько высоким, чтобы $\forall x_j \in X, u_j = n, n = \text{card}(X)$.

Таким образом, если некоторая мода является внутренней точкой подмножества, то она является локальным максимумом функции $\mu_B(x_i)$.

Авторская версия алгоритма состоит из двух процедур, причем первая процедура, в свою очередь, состоит из двух частей: первая часть разбивает дискретное нечеткое множество B точек, подлежащих классификации, на симметричные нечеткие подмножества, а вторая часть отыскивает локальные максимумы мультимодальной функции принадлежности нечеткого множества B . Вторая процедура разбивает нечеткое множество B на унимодальные нечеткие множества, используя в качестве основы полученную в результате работы первой процедуры информацию о числе, взаимном расположении и значениях функции принадлежности точек $\tau^l, l = 1, \dots, c$ — локальных максимумов функции принадлежности $\mu_B(x_i)$. В изложенном ниже варианте алгоритм представлен в виде единой кластер-процедуры.

Параметры алгоритма:

Входными данными является матрица $X_{\text{мат}} = [x_i^l]$; параметром алгоритма — порог различия объектов $\hat{d}$;

Схема алгоритма:

1 Строятся следующие последовательности:

1.1. последовательность $A = \{y_1, y_2, \dots, y_n\}$ точек $x_i, i = 1, \dots, n$ исследуемой совокупности, $x_i \in B$, упорядоченных в соответствии со значениями их функции принадлежности, так что $y_i = \arg \max_i \mu_B(x_i), y_n = \arg \min_i \mu_B(x_i)$ и $y_i \leq y_j$ при $\mu_B(x_i) \geq \mu_B(x_j)$, где $\mu_B(x_i)$ и $\mu_B(x_j)$ — функции принадлежности точек x_i и x_j соответственно;

1.2. последовательность $A_{(1)} = \{y_1^{(1)}, \dots, y_n^{(1)}\}$, представляющая собой упорядоченную в соответствии с их расстояниями до точки $y_1^{(1)} = y_1$, то есть $d(y_1^{(1)}, y_j^{(1)}) \leq d(y_1^{(1)}, y_i^{(1)})$ при $j \leq i$, последовательность точек — «кандидатов» в группу $\tilde{A}^1$, то есть группу, для которой точка $y_1^{(1)} = y_1 = \hat{\tau}^1$ будет являться модой, так что

первым «кандидатом» в группу $\tilde{A}^1$ будет точка $y_2^{(1)}$, следующей точкой — «кандидатом» в группу $\tilde{A}^1$ будет точка $y_3^{(1)}$, и так далее;

2 Если $y_i = y_i^{(1)}$ для $i = 2, 3, \dots, r-1$ и $y_r \neq y_r^{(1)}$, то $y_i, i = 1, 2, \dots, r-1$ распределяются в группу $\tilde{A}^1$, а точка $y_1^{(2)} = y_r = \hat{t}^2$ рассматривается как мода второй группы, так что формируется последовательность $A_{(2)} = \{y_1^{(2)}, \dots, y_p^{(2)}\}$, в которой точки упорядочиваются по расстоянию до точки $y_1^{(2)} = y_r = \hat{t}^2$, так что $d(y_1^{(2)}, y_j^{(2)}) \leq d(y_1^{(2)}, y_t^{(2)})$ при $j \leq t$; подобный процесс выполняется для любой последовательности $A_{(f)} = \{y_1^{(f)}, \dots, y_i^{(f)}\}$, так что если $y_1^{(f)} = \hat{t}^f$ — мода группы $\tilde{A}^f$, то первым «кандидатом» в группу $\tilde{A}^f$ будет точка $y_2^{(f)}$;

3 Производится формирование групп в соответствии со следующим правилом: пусть к некоторому моменту построено G групп, то есть сконструированы последовательности $A_{(g)}, g = 1, \dots, G$ точек — «кандидатов» $y_k^{(g)}$, и пусть некоторую точку $y_q \in A$ требуется приписать некоторой группе $\tilde{A}^g, g = 1, \dots, G$, тогда как все предшествующие ей точки $y_i \in A, i < q$ расклассифицированы, тогда оказываются возможными следующие ситуации:

3.1. если $y_q = y_k^{(g)}$ и $y_q = y_k^{(m)}, m = 1, \dots, G, m \neq g$, то точка y_q включается в формируемую группу $\tilde{A}^g$;

3.2. если $y_q = y_k^{(g)}$ для некоторого $g \in L$, где L — множество индексов, представляющих те группы, точки — «кандидаты» которых идентичны y_q , то точка y_q включается в ту группу, к которой приписан ее ближайший сосед, имеющий наибольшее значение функции принадлежности;

3.3. если $y_q \neq y_k^{(g)}$ для $g = 1, \dots, G$, то формируется новая группа, для которой точка y_q является модой;

Процесс продолжается до исчерпания последовательности $A = \{y_1, y_2, \dots, y_n\}$;

4 Для каждой группы $\tilde{A}^g, g = 1, \dots, G$ и соответствующей ей моды $\hat{t}^g$ находится минимальное расстояние $R_g = d(\hat{t}^g, x_{p(g)}) = \min_{x_k \in (B - \tilde{A}^g)} [d(\hat{t}^g, x_k)]$

до точки $x_{p(g)} \in B$, являющейся ближайшей к группе $\tilde{A}^g$ и лежащей вне ее: $x_{p(g)} \notin \tilde{A}^g$; точка $\hat{t}^g = \tau^g$ будет локальным максимумом, если множество $B_{R_g} = \{x_i \mid d(x_{p(g)}, x_i) < R_g\}$ включает в себя точки группы $\tilde{A}^g$, в противном случае точка $\hat{t}^g$ является не локальным максимумом, а граничной точкой группы $\tilde{A}^g$; данная процедура применяется ко всем модам $\hat{t}^g \in \tilde{A}^g, g=1, \dots, G$, так что формируется последовательность локальных максимумов $\tau^l, l=1, \dots, c$, упорядоченная в соответствии со следующим правилом: для локальных максимумов τ^i и $\tau^j, i, j=1, \dots, c$ при $\mu_B(\tau^i) > \mu_B(\tau^j)$ локальные максимумы τ^i и τ^j расположены так, что $i \leq j$;

- 5 Ко всем точкам, за исключением локальных максимумов $\tau^l, l=1, \dots, c$, порождающих кластеры $A^l, l=1, \dots, c$, применяется следующее правило: точка $x_i \in B$, расположенная в последовательности A в j -й позиции, то есть $x_i = y_j$, распределяется в ту группу A^l , в которой находится ее ближайший сосед с более высоким значением функции принадлежности, так что строится разбиение $P = \{A^1, \dots, A^c\}$ и алгоритм прекращает работу.

Приведенная версия процедуры основана на предположении, что при $i \neq j$ выполняется неравенство $\mu_B(x_i) \neq \mu_B(x_j)$. Вместе с тем, возможной оказывается также ситуация, когда значение функции принадлежности оказывается одинаковым для нескольких точек, то есть $\mu_B(x_i) = \mu_B(x_j), i \neq j, i, j \in [1, n]$. В таком случае проблема решается внесением изменения на шаге 3 процедуры и перестановкой элементов в последовательности A следующим образом: при сформированных G группах некоторая точка $y_q \in A$ подлежит включению в некоторую группу $\tilde{A}^g, g=1, \dots, G$. Если $y_q \neq y_k^{(g)}$ для $g=1, \dots, G$ и если $\mu_B(y_{q+1}) = \mu_B(y_q)$, то необходимо проверить выполнение равенства $y_{q+1} = y_k^{(g)}, g=1, \dots, G$. Новая группа с модой y_q образуется только в том случае, если ни одна из точек $y_{q+1}, y_{q+2}, \dots$, имеющих такое же значение функции принадлежности, как и точка y_q , не идентична с точкой $y_k^{(g)}, g=1, \dots, G$.

Может оказаться, что наибольшее значение функции принадлежности в некоторой группе $\tilde{A}^g$ получают несколько точек, поэтому

сложно определить, какая из них является локальным максимумом. В качестве метода, позволяющего преодолеть подобное затруднение, может быть использовано на шаге 4 приведенной выше процедуры следующее правило: пусть $\tilde{A}_t^s$ — подмножество точек из $\tilde{A}^s$, имеющих, как и точка $\hat{t}^s$, максимальное значение функции принадлежности; каждая из точек подмножества $\tilde{A}_t^s$ проверяется как мода группы $\tilde{A}^s$, и если хотя бы одна из них является пограничной для группы $\tilde{A}^s$, то точка $\hat{t}^s$ не рассматривается как локальный максимум.

Следует также указать, что в изложении схемы алгоритма, представленной в [22, с.53-56], не описывается шаг, соответствующий второй процедуре авторской версии и, соответственно, шагу 5 версии, представленной выше, где группы $\tilde{A}^s, g=1, \dots, G$ нечеткого множества B порождают унимодальные нечеткие множества $A^l, l=1, \dots, c$. Вместо этого предлагается следующая процедура построения нечетких кластеров $A^l, l=1, \dots, c$: пусть $\tau^l, l=1, \dots, c$ — сформированная последовательность локальных максимумов; если на множестве $X = \{x_1, \dots, x_n\}$ задана нечеткая толерантность T , так что $\mu_T(x_i, x_j)$ — степень сходства элементов $x_i, x_j \in X, i, j=1, \dots, n$, то степень принадлежности μ_{ii} элемента x_i кластеру A^l может быть вычислена по формуле $\mu_{ii} = \mu_T(x_i, x_i) / \left(\sum_{j=1}^c \mu_T(x_i, \tau^j) \right)$. Многочисленные вычислительные эксперименты — для 1000 сгенерированных объектов — показали достаточно высокую скорость вычислений при небольших затратах памяти, а получаемые кластеры могут иметь как эллиптическую форму, так и форму окружности.

3.1.2.2. Алгоритм Тамуры — Хигути — Танаки [163] использует (max-min)-транзитивное замыкание нечеткой толерантности T , описывающей исходные данные в виде матрицы близости $r_{n \times n} = [\mu_T(x_i, x_j)], i=1, \dots, n, j=1, \dots, n$, и состоит из трех шагов.

Основные понятия и определения

Пусть $X = \{x_1, \dots, x_n\}$ — множество классифицируемых объектов, на котором задана нечеткая толерантность T в виде матрицы $r_{n \times n} = [\mu_T(x_i, x_j)]$ близости объектов (1.3).

Параметры алгоритма:

c — число кластеров в искомом разбиении P^* ;

Схема алгоритма:

- 1 Строится транзитивное замыкание $\tilde{T}$ исходного отношения нечеткой толерантности T в соответствии с формулой (2.40);
- 2 Вычисляются α -уровни, $\alpha \in [0,1]$ полученного нечеткого отношения эквивалентности $\tilde{T}$ и для каждого $\alpha, \alpha \in [0,1]$ выделяются кластеры $A^l, l \in [1, n]$ в соответствии с правилом: если $\mu_{\tilde{T}}(x_i, x_j) \geq \alpha$, для некоторых $x_i, x_j \in X$, то объекты $x_i, x_j \in X$ принадлежат кластеру A^l ;
- 3 Из полученной иерархии разбиений выбирается разбиение на $c, 2 \leq c \leq n$ классов $P^* = \{A^1, \dots, A^c\}$, соответствующее некоторому порогу $\alpha, \alpha \in [0,1]$, и алгоритм прекращает работу.

Таким образом, объекты оказываются однозначно распределенными по четким классам, представляющим собой множества уровня α . Результат представляется в виде матрицы разбиения $P_{\text{сх}} = [\mu_{ii}]$, где

$$\mu_{ii} = \begin{cases} 1, & x_i \in A^l \\ 0, & x_i \notin A^l \end{cases}, l = 1, \dots, c, i = 1, \dots, n, \text{ или компонент графа [163, с.65].}$$

3.1.2.3. Алгоритм Кутюрье — Фьолео [71] использует понятие так называемого максимально внутренне устойчивого множества (stable set internally maximal). Особенностью метода является возможность пересечения нечетких кластеров.

Основные понятия и определения

Пусть $X = \{x_1, \dots, x_n\}$ — множество классифицируемых объектов, на котором задано нечеткое отношение обычного несходства I в виде соответствующей матрицы $d_{\text{нх}} = [\mu_I(x_i, x_j)], i = 1, \dots, n, j = 1, \dots, n$. Пусть для некоторого $\alpha \in [0,1]$ четкое отношение I_α является α -уровнем нечеткого отношения обычного несходства I и A' — некоторое подмножество исследуемой совокупности объектов, $A' \subset X$. Пусть $I_\alpha(A') = \{x_j \in X : \forall x_i \in A', (x_i, x_j) \in I_\alpha\}$ обозначает множество объектов, наиболее отличающихся по крайней мере от некоторого одного элемента $x_j \in X$. Тогда подмножество A' будет именоваться максимально внутренне устойчивым множеством, если выполняются условие $A' \cap I_\alpha(A') = \emptyset$ и условие $A' \cup I_\alpha(A') = X$. При вычислении максимально внутренне устойчивых множеств рассматриваются три варианта:

- 1) нет элемента $x_j \in A^i$, такого, что для некоторого x_i выполняется $(x_i, x_j) \in I_\alpha$: максимально внутренне устойчивое множество A^i замещается множеством $A^i \cup \{x_i\}$;
- 2) для всех элементов $x_j \in A^i$ выполняется условие $(x_i, x_j) \in I_\alpha$: добавляется новое максимально внутренне устойчивое множество $\{x_i\}$;
- 3) только для некоторых элементов $x_j \in A^i$ выполняется условие $(x_i, x_j) \in I_\alpha$, так что максимально внутренне устойчивое множество A^i разбивается на два подмножества, A_1^i и A_2^i , где A_1^i — подмножество элементов A^i , подобных элементу x_i , а A_2^i — подмножество элементов A^i , отличных от элемента x_i : создается новое максимально внутренне устойчивое множество $A_1^i \cup \{x_i\}$.

Новое множество максимально внутренне устойчивых множеств упрощено: все максимально внутренне устойчивые множества, включенные в другие максимально внутренне устойчивые множества, выбывают из рассмотрения.

Кластеры представляют собой максимально внутренне устойчивые множества, удовлетворяющие двум условиям:

- 1) условию представительства, в соответствии с которым любое максимально внутренне устойчивое множество с небольшим числом элементов не рассматривается; максимально внутренне устойчивое множество A^i является кластером, если $\text{card}(A^i) \geq u$, где u — порог, задаваемый исследователем;
- 2) условию делимости, в соответствии с которым для любых двух максимально внутренне устойчивых множеств число элементов в области их пересечения не должно превышать пороговое значение w , задаваемое исследователем.

Таким образом, множество кластеров $C = \{A^1, \dots, A^c\}$ должно удовлетворять следующему условию:

$$C = \{\forall A^i \in \mathfrak{Z}, i = 1, \dots, c : \text{card}(A^i) \geq u, \\ \forall A^i, A^f \in \mathfrak{Z}, i, f = 1, \dots, c, i \neq f, \text{card}(A^i \cap A^f) \leq w\}, \quad (3.3)$$

где символ $\mathfrak{Z}$ обозначает множество максимально внутренне устойчивых множеств.

Параметры алгоритма:

α — порог различия элементов, объединяемых в нечеткие кластеры, $\alpha \in (0,1]$;

u — минимальное число элементов в нечетком кластере $A^l, l = 1, \dots, c$;

w — максимальное число элементов в области пересечения любых двух нечетких кластеров;

Схема алгоритма:

- 1 Для некоторого задаваемого исследователем порога различия $\alpha, \alpha \in (0,1]$ строится I_α — α -уровень нечеткого отношения несходства I в соответствии с условием

$$\mu_{I_\alpha}(x_i, x_j) = \begin{cases} 0, \mu_l(x_i, x_j) \leq \alpha \\ 1, \mu_l(x_i, x_j) > \alpha \end{cases}, i, j = 1, \dots, n;$$

- 2 Вычисляются максимально внутренне устойчивые множества для построенного четкого отношения несходства I_α ;
- 3 В соответствии с критерием (3.3) для заданных исследователем параметров u и w из множества $\mathfrak{Z}$ максимально внутренне устойчивых множеств выделяются кластеры $A^l, l = 1, \dots, c$;
- 4 Для каждого кластера $A^l, l = 1, \dots, c$ определяется точка $\tau^l, l = 1, \dots, c$, являющаяся его центром, и строится функция принадлежности μ_{I_l} , которая удовлетворяет следующему условию:

$$\mu_{I_l} = \begin{cases} 1, x_i = \tau^l \\ 0, \mu_l(x_i, \tau^l) = \max \mu_l(x_i, x_j) \end{cases}, i, j = 1, \dots, n, l = 1, \dots, c.$$

- 5 Строится матрица $C_{\text{сх}} = [\mu_{I_l}]$ нечеткого покрытия $C = \{A^1, \dots, A^c\}$, элементами которой являются степени схождения $\mu_{I_l} = \mu(x_i, \tau^l), i = 1, \dots, n, l = 1, \dots, c$ элементов множества $X = \{x_1, \dots, x_n\}$ с центрами $\tau^1, \dots, \tau^c$ кластеров $A^1, \dots, A^c$ в соответствии со следующей формулой:

$$\mu_{I_l} = 1 - \frac{\mu_l(x_i, \tau^l)}{\max \mu_l(x_i, x_j)}, i, j = 1, \dots, n, l = 1, \dots, c, \text{ и алгоритм прекращает}$$

работу.

Алгоритм был применен [71, с.40-44] для анализа состояния 165 французских фирм в 1995 финансовом году при различных значениях входных параметров и в сравнении с алгоритмом Беждека — Данна показал хорошие результаты.

3.1.2.4. Алгоритм Берштейна — Дзюбы [11] в качестве критерия оптимизации разбиения вершин исходного нечеткого графа на два класса использует наибольшую степень двудольности.

Основные понятия и определения

Пусть $G = (X, T)$ — нечеткий граф, где $X = \{x_1, \dots, x_n\}$ — множество вершин, а $T = \{\mu_{ij} / t_{ij}\}, i, j = 1, \dots, n$ — нечеткое множество ребер нечеткого графа G , где $\mu_{ij} = \mu_T(x_i, x_j)$ — функция принадлежности ребра $t_{ij} = \langle x_i, x_j \rangle$, так что $\mu_T : X \times X \rightarrow [0, 1]$.

Степень двудольности нечеткого графа $G = (X, T)$ определяется в соответствии с выражением

$$\Delta = \max_l \delta_l, \quad (3.4)$$

где $l = 1, \dots, c$, символом δ_l обозначается степень двудольности l -й двудольной части нечеткого графа G , а символом c — число различных максимальных двудольных частей данного нечеткого графа G .

Для определения степени двудольности δ_l может быть использован любой из двух следующих способов.

Первый способ определения степени двудольности заключается в том, что любую часть $G' = (X' \cup X'', T')$, где $X' \cup X'' = X$ и $G' \subseteq T'$, можно рассматривать как нечеткую двудольную часть графа G со степенью двудольности, вычисляемой в соответствии с формулой

$$\delta_l = 1 - \max(\mu_I, \mu_{II}), l = 1, \dots, c, \quad (3.5)$$

где значения μ_I и μ_{II} определяются по формулам

$$\mu_I = \max_{t_{ij} \in I} \mu_{ij}, I = \{\langle x_i, x_j \rangle, x_i, x_j \in X'\}, \quad (3.6)$$

$$\mu_{II} = \max_{t_{ij} \in II} \mu_{ij}, II = \{\langle x_i, x_j \rangle, x_i, x_j \in X''\}. \quad (3.7)$$

В дальнейшем, для упрощения обозначений, ребра, являющиеся элементами множества I , будут обозначаться символом t^I , а ребра, являющиеся элементами множества II , — символом t^{II} .

Второй способ определения степени двудольности состоит в ее вычислении по следующей формуле:

$$\delta_l = 1 - \frac{\sum \mu(t^I) + \sum \mu(t^{II})}{M(X') + M(X'')}, \quad (3.8)$$

где, в свою очередь, $\mu(t^I)$ — значение функции принадлежности некоторого ребра $t^I, t^I \in I$; $\mu(t^{II})$ — значение функции принадлежности

некоторого ребра t'' , $t'' \in H$, $M(X') = \frac{\text{card}(X')(\text{card}(X') - 1)}{2}$ — максимально возможное число ребер между вершинами множества X' , $M(X'') = \frac{\text{card}(X'')(\text{card}(X'') - 1)}{2}$ — максимально возможное число ребер между вершинами множества X'' , а символами $\text{card}(X')$, $\text{card}(X'')$ обозначено число вершин во множестве X' и множестве X'' соответственно. В дальнейшем степень двудольности, вычисленная в соответствии с первым способом, то есть по формуле (3.5), будет обозначаться символом δ_i^1 , а если для этой цели применяется второй способ, то есть формула (3.8) — соответственно, символом δ_i^2 , поскольку, в общем, $\delta_i^1 \neq \delta_i^2$. Таким образом, представляется очевидным, что если во множествах вершин X' и X'' графа, полученного в результате выделения двудольной части, не существует ни одной дуги, то степень двудольности такого графа равна 1, причем независимо от способа вычисления величины δ_i .

Максимальная двудольная часть графа G' представляет собой двудольную часть, не являющуюся частью никакой другой. Если число вершин в нечетком графе G' является нечетным, то в общем случае число максимальных двудольных частей, которые можно выделить в данном максимальном нечетком графе, может быть вычислено по формуле

$$c = C_n^1 + C_n^2 + \dots + C_n^{\lfloor n/2 \rfloor} = \sum_{f=1}^{\lfloor n/2 \rfloor} C_n^f, \quad (3.9)$$

где символом $\lfloor n/2 \rfloor$ обозначается округление величины $n/2$ в меньшую сторону, а символ C_n^f традиционно обозначает число сочетаний:

$$C_n^f = \frac{n!}{f!(n-f)!}. \quad (3.10)$$

Если же число вершин в нечетком графе G' является четным, то число максимальных двудольных частей, которые можно выделить в данном максимальном нечетком графе, вычисляется в соответствии с формулой

$$c = C_n^1 + C_n^2 + \dots + C_n^{n/2-1} / 2 = \sum_{f=1}^{\lfloor n/2 \rfloor - 1} C_n^f + C_n^{n/2} / 2. \quad (3.11)$$

Критерием оптимизации разбиения вершин исходного нечеткого графа на два класса является наибольшая степень двудольности δ_i^1 или δ_i^2 .

Параметры алгоритма:

Нечеткий граф задается в виде матрицы $r_{n \times n} = [\mu_T(x_i, x_j)]$ отношения нечеткой толерантности. Ребра $\mu_{ij}/t_{ij}, i=1, \dots, n$ в процессе выполнения алгоритма не рассматриваются. Входные параметры как таковые отсутствуют.

Схема алгоритма:

- 1 В нечетком графе G' ранжируются все ребра $t_{ij}, i, j=1, \dots, n, i \neq j$ в порядке убывания их степеней принадлежности μ_{ij} , так что формируется последовательность $t^{(1)}, t^{(2)}, \dots$;
- 2 Рассматривается первое ребро $t^{(1)}$ с максимальной степенью принадлежности μ_{ij} из последовательности $t^{(1)}, t^{(2)}, \dots$, и вершина x_i , инцидентная этому ребру, помещается в первую долю, то есть список X' , а вершина x_j — в другую долю, то есть список X'' ;
- 3 Для остальных ребер из последовательности $t^{(1)}, t^{(2)}, \dots$:
 - 3.1.если $x_i \in X'$, то вершина x_j помещается в список X'' ;
 - 3.2.если $x_i \in X''$, то вершина x_j помещается в список X' ;
 - 3.3.если $x_j \in X'$, то вершина x_i помещается в список X'' ;
 - 3.4.если $x_j \in X''$, то вершина x_i помещается в список X' ;
 - 3.5.если $x_i \in X', x_j \in X''$ или $x_i \in X'', x_j \in X'$, то ребро t_{ij} прибавляется к нечеткой двудольной части;
 - 3.6.если $x_i \notin X'$ и $x_j \notin X''$ или $x_i \notin X''$ и $x_j \notin X'$, то оказываются возможны два варианта:
 - 3.6.1. вершина x_i помещается в список X' , вершина x_j помещается в список X'' ;
 - 3.6.2. вершина x_i помещается в список X'' , вершина x_j помещается в список X' ;
- 4 Если были просмотрены все ребра, вычислить степени двудольности δ_i^1 и/или δ_i^2 по соответствующим формулам, и алгоритм прекращает работу; в противном случае осуществляется переход на шаг 3.

Частными случаями задачи являются ситуации, когда оказывается необходимым выделить для исходного графа с четным числом вершин такую максимальную нечеткую двудольную часть с наибольшей степенью двудольности, которая содержит одинаковое число вершин в каждой из долей, а для графа с нечетным числом вершин необходимо выделить такую максимальную нечеткую двудольную часть с наибольшей степенью двудольности, которая содержит в одной доле $\lfloor n/2 \rfloor$ вершин и в другой доле $\lfloor n/2 \rfloor + 1$ вершин. В таких случаях в шаг 3 вышеприведенной версии алгоритма следует включить проверку дополнительного условия в одной из следующих двух формулировок:

- 1) если число вершин n в исходном графе G' — четное и $\text{card}(X') \geq n/2$ или число вершин n в исходном графе G' — нечетное и $\text{card}(X') \geq \lfloor n/2 \rfloor + 1$, то все остальные, не просмотренные ранее вершины помещаются в список X'' ;
- 2) если число вершин n в исходном графе G' — четное и $\text{card}(X'') \geq n/2$ или число вершин n в исходном графе G' — нечетное и $\text{card}(X'') \geq \lfloor n/2 \rfloor + 1$, то все остальные, не просмотренные ранее вершины помещаются в список X' .

Таким образом, после упорядочивания ребер на шаге 1 сложность алгоритма зависит от числа несмежных ребер исходного графа, которые находятся в последовательности $t^{(1)}, t^{(2)}, \dots$ рядом друг с другом [11, с.119-120].

Необходимо также еще раз подчеркнуть, что данный алгоритм разбивает множество объектов $X = \{x_1, \dots, x_n\}$, представленных в виде вершин графа, на два класса. Если же при решении задачи классификации множество объектов $X = \{x_1, \dots, x_n\}$ необходимо разбить на большее число групп, то сначала выделяется двудольная структура, после чего каждый из выделенных классов снова разбивается на два подкласса и так далее до тех пор, пока классифицируемое множество $X = \{x_1, \dots, x_n\}$ не будет разделено на заданное число классов [11, с.121].

3.1.2.5. Другие алгоритмы, представляющие эвристическое направление нечеткого подхода в кластерном анализе, также сравнительно немногочисленны. К примеру, в работе [123] описывается метод, использующий, как и в алгоритме Тамуры — Хигути — Танаки, операцию транзитивного замыкания, а в исследовании И. З. Батырши-

на [47] рассматривается процедура кластеризации, исходными данными для которой также является нечеткое отношение сходства.

В работе [178] описывается человеко-машинный подход к решению нечеткой модификации задачи кластерного анализа. Сущность предлагаемого подхода состоит в том, что нечеткие кластеры, являющиеся нечеткими множествами уровня α , в свою очередь, порожденными некоторой нечеткой толерантностью T , матрица которой является матрицей исходных данных, образуют так называемое представление $R_z^\alpha(X) = \{A_\alpha^l \mid l = \overline{1, c}, c \leq n, \alpha \in (0, 1]\}$ исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$ для некоторого уровня $\alpha, \alpha \in (0, 1]$, если вы-

полняется условие $\sum_{l=1}^c \mu_{li} > 0, \forall x_i \in X$ и число нечетких кластеров в представлении оказывается не меньшим, чем два, так что решение задачи классификации подразумевает выбор некоторого единственного представления $R^*(X)$ из множества различных представлений для различных значений $\alpha, \alpha \in (0, 1]$. Значение α , таким образом, представляет собой порог сходства элементов, объединяемых в некоторый нечеткий кластер. Метод оправдан с гносеологической точки зрения, позволяет выделять кластеры сложной формы и классифицировать объекты, когда матрица близости классифицируемых объектов представляет собой матрицу любой нечеткой толерантности $T_b, b = 0, 1, 2, 3$; более того, поскольку нечеткие кластеры представляются как нечеткие множества уровня, то допускается их пересечение [17]. Вместе с тем, применение данного подхода вызывает затруднения обработки данных при большом объеме исследуемой совокупности; кроме того, выбор термина «представление» (representation), используемого для определения искомой структуры, является неудачным, поскольку традиционно в задачах кластер-анализа данный термин имеет отличный от используемого в [178] смысл.

3.2. Оптимизационные методы

3.2.1. Нечеткая модификация экстремальной постановки задачи автоматической классификации и функционалы качества нечеткого разбиения

Как отмечал К. Яюга [106, с.410], наиболее общим и распространенным подходом к решению задачи классификации в условиях нестохастической неопределенности является оптимизационный под-

ход. В подобных методах нечеткая кластеризация понимается как разбиение множества данных на совокупность его нечетких множеств, так что в качестве входного параметра в существующих нечетких методах оптимизационного направления автоматической классификации задается число нечетких кластеров, причем при данном подходе под нечетким кластером может пониматься любое нечеткое множество, определенное на универсуме. Нечеткие множества $A^l, l=1, \dots, c$, определенные на универсуме $X = \{x_1, \dots, x_n\}$, с соответствующими функциями принадлежности $\mu_{li}, l=1, \dots, c, i=1, \dots, n$, образуют нечеткое c -разбиение $P = \{A^1, \dots, A^c\}$ в смысле Э. Г. Распини, если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^c \mu_{li} = 1$ и нечеткая модифика-

ция задачи автоматической классификации в экстремальной постановке заключается в нахождении экстремума, как правило, минимума, некоторого функционала $Q(P)$ на множестве всех нечетких разбиений, что также описывается формулой (1.5) с учетом того обстоятельства, что Π — множество всех возможных нечетких разбиений P исходного множества объектов X . Нечеткие разбиения подробно исследуются в работах [51], [54], [57]. Необходимо также указать, что условие, в соответствии с которым нечеткие множества, определенные на универсуме $X = \{x_1, \dots, x_n\}$, образуют нечеткое разбиение, как отмечал Л. А. Заде [25, с.230], является «достаточно сильным предположением».

Если исходные данные представлены матрицей расстояний $d_{n \times n} = [\mu_{ij}(x_i, x_j)], i=1, \dots, n, j=1, \dots, n$, то функционал качества разбиения в общем виде выглядит следующим образом [148, с.240]:

$$Q'(P) = \sum_{i=1}^c \sum_{l=1}^n \sum_{j=1}^n f(p(x_i), v_{li}) g(p(x_j), \mu_{lj}) d(x_i, x_j), \quad (3.12)$$

где f, g — некоторые функционалы; $p(x_i)$ — априорный вес для каждой точки $x_i, \sum_{i=1}^n p(x_i) = 1$; μ_{li} — функция принадлежности элемента $x_i \in X$ некоторому нечеткому кластеру $A^l \in \{A^1, \dots, A^c\}$; v_{lj} — некоторая вспомогательная функция принадлежности, $d(x_i, x_j)$ — функция, определяющая расстояние между объектами x_i и x_j , как правило, представляющая собой элемент матрицы $d_{n \times n}$, и где в качестве ограничений выступают условия

$$\mu_{ii} \geq 0, \sum_{i=1}^c \mu_{ii} = 1, v_{ii} \geq 0, \sum_{i=1}^n v_{ii} > 0, i = 1, \dots, n, l = 1, \dots, c, \quad (3.13)$$

так что нахождение классификации состоит в решении оптимизационной задачи

$$Q'(P) \rightarrow \min \quad (3.14)$$

с соответствующими ограничениями.

При анализе функционала качества разбиения, предложенного Э. Г. Распини [149] в 1969 году и являющегося первым функционалом, сформулированным в терминах нечетких множеств, необходимо, следуя за автором критерия качества, кратко рассмотреть общую математическую модель задачи, поскольку, как отмечал М. П. Уиндхем, «к сожалению, критерий, на котором основан алгоритм Распини, труден для интерпретации, а численная процедура довольно сложна» [186, с.159]. Подобная точка зрения высказывается также Дж. Беждеком [59, с.58]. Более того, подобное рассмотрение оказывается нелишним для исследования эволюции нечетких методов кластер-анализа.

Пусть, как и ранее, $X = \{x_1, \dots, x_n\}$ — множество n объектов и $p(x)$ — функция распределения вероятностей, определенная в X , а символом P обозначается множество кластеров $\{A^1, \dots, A^c\}$, являющихся результатом группировки объектов исходной совокупности X , тогда функция $r: X \rightarrow P$, в свою очередь, будет именоваться группообразующей, если она удовлетворяет следующим условиям: для каждого $A^i \in P$ существуют функции

$$T_i: X \times X \rightarrow R, \quad (3.15)$$

$$V_i: P \times P \rightarrow R, \quad (3.16)$$

где $X \times X$ и $P \times P$ — декартовы произведения множеств X и P самих на себя соответственно и R — множество вещественных чисел, таких, что для всех пар $(x_i, x_j) \in X \times X$ выполняется условие

$$T_i(x_i, x_j) = V_i(r(x_i), r(x_j)). \quad (3.17)$$

С содержательной точки зрения, группообразующая функция r устанавливает соответствие между элементами исследуемой совокупности $X = \{x_1, \dots, x_n\}$ и элементами множества кластеров $P = \{A^1, \dots, A^c\}$, а $\{T_i, V_i\}$ — множество ограничений, выполнение которых необходимо для сохранения при преобразовании определенных свойств элементов исследуемой совокупности X .

Если же группообразующая функция r не только удовлетворяет множеству ограничений $\{T_i, V_i\}$, но и доставляет экстремум некоторой целевой функции, то в этом случае полученное разбиение $P^* = \{A^1, \dots, A^c\}$ оказывается в некотором смысле наилучшим на множестве всех возможных разбиений, порождаемых группообразующей функцией r . К примеру, группообразующая функция r может доставлять минимум некоторой целевой функции:

$$Q: P \rightarrow R, Q(r) = \min. \quad (3.18)$$

Однако на практике ограничения являются столь жесткими, что группирующую функцию указать невозможно, для чего приходится вводить различные допущения. Если смягчить все ограничения, то в общем случае целевая функция примет следующий вид:

$$Q(r) = \sum_{i=1}^c W_i^2 \iint (V_i(r(x_i), r(x_j)) - T_i(x_i, x_j)) dp(x_i) dp(x_j), \quad (3.19)$$

где W_i — весовые функции на общие ограничения.

Функционал $Q'(P)$ является, в общем, частным случаем соотношения (3.19). При построении функционала Э. Г. Распини использовал два условия для выбора вида группообразующей функции. Если для каждой точки x_i допускается нечеткий кластер A^i , такой, что $\mu_{ii} = \max_{1 \leq l \leq c} \mu_{li}$, то в соответствии с первым условием группообразующая функция r должна минимизировать соотношение

$$Q'_{(1)} = \sum_{i=1}^n \left(\frac{p(x_i) \frac{\sum_{j=1}^n p(x_j) \mu_{ij} d(x_i, x_j)}{i-1}}{\sum_{j=1}^n p(x_j) d(x_i, x_j)} \right)^2, \quad (3.20)$$

выражающее среднюю плотность точек в том нечетком кластере, к которому они относятся с наибольшим значением функции принадлежности.

Если в формуле (3.12) положить $v_{ii} = \mu_{ii}$ [148, с.244-245] и будет иметь место

$$f(p(x_i), v_{ii}) g(p(x_j), \mu_{ij}) = p(x_i) \mu_{ii} p(x_j) \mu_{ij} \frac{\sum_{j=1}^n p(x_j) \mu_{ij}}{\sum_{j=1}^n p(x_j) d(x_i, x_j)}, \quad (3.21)$$

и

$$\sum_{l=1}^c \mu_{il} = 1, \quad (3.22)$$

тогда в соответствии со вторым условием, группообразующая функция r должна быть выбрана таким образом, чтобы соотношение

$$Q_{1(2)}^i = \sum_{i=1}^n p(x_i) \sum_{l=1}^c \left(\sum_{j=1}^n p(x_j) \mu_{lj} \right) \left(\frac{\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)}{\sum_{j=1}^n p(x_j) d(x_i, x_j)} \right) \mu_{li} \quad (3.23)$$

также достигало минимума. Минимизация соотношения (3.23) мини-

мизирует в среднем произведение $\mu_{li} \cdot \frac{\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)}{\sum_{j=1}^n p(x_j) \mu_{lj}}$, так что наи-

большие средние плотности будут соответствовать наименьшим степеням принадлежности.

В приведенных выше соотношениях выражение $\sum_{j=1}^n p(x_j) d(x_i, x_j)$ представляет собой ни что иное, как среднюю плотность точек вокруг точки x_i , а выражение $\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)$ в соотношении (3.23) полу-

чается из произведения $\sum_{j=1}^n p(x_j) \mu_{lj} \cdot \frac{\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)}{\sum_{j=1}^n p(x_j) \mu_{lj}}$, где, в свою

очередь, $\sum_{j=1}^n p(x_j) \mu_{lj}$ показывает относительный размер нечеткого кла-

стера A^l , а выражение $\frac{\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)}{\sum_{j=1}^n p(x_j) \mu_{lj}}$ представляет собой сред-

нюю плотность точек вокруг точки x_i в нечетком кластере A^l . Аналогичные рассуждения можно провести относительно выражения $\sum_{i=1}^n p(x_i) \mu_{li} d(x_i, x_i)$ в соотношении (3.20). Соотношения (3.20) и (3.23) рассматривались Э. Г. Распини в качестве предварительного этапа при

построении функционала, и поиск оптимума любого из них также может быть использован как метод кластеризации.

Таким образом, с учетом ограничений на группообразующую функцию

$$T(x_i, x_j) = F(d(x_i, x_j)), \quad (3.24)$$

где $F(d(x_i, x_j))$ — неотрицательная неубывающая функция, такая, что $F(0) = 0$, и

$$V(r(x_i), r(x_j)) = \sum_{i=1}^c (\mu_{ii} - \mu_{ij})^2, \quad (3.25)$$

которые, исходя из формулы (3.19), можно объединить в одно условие оптимальности [150, с.343]

$$Q(r) = \sum_{i=1}^n \sum_{j=1}^n p(x_i) p(x_j) (V(r(x_i), r(x_j)) - T(x_i, x_j))^2, \quad (3.26)$$

функционал Э. Г. Распини примет вид

$$Q'_1(P) = \sum_{i=1}^n \sum_{j=1}^n p(x_i) p(x_j) \left(\sum_{i=1}^c (\mu_{ii} - \mu_{ij})^2 - F(d(x_i, x_j)) \right)^2. \quad (3.27)$$

Величина, определяемая формулой (3.25), представляет собой измеритель общей близости объектов на множестве $P = \{A^1, \dots, A^c\}$ и равна единице, если объекты со степенью принадлежности, равной единице, попадают в два разных класса, а если со степенью принадлежности, равной единице, объекты попадают в один класс, то эта величина равна, соответственно, нулю. Таким образом, величину $V(r(x_i), r(x_j))$ можно трактовать и как своего рода «контекстную» меру близости объектов. В свою очередь, величина $F(d(x_i, x_j))$ характеризует прямую близость объектов, причем очевидно, что чем ближе объекты по величине $d(x_i, x_j)$, тем ближе они и по величине $F(d(x_i, x_j))$. В качестве функции, характеризующей прямую близость объектов, может рассматриваться, например, соотношение $F(d(x_i, x_j)) = \Psi d^2(x_i, x_j)$, где $\Psi = \text{const}, \Psi > 0$, при которой, как отмечал Э. Г. Распини [150], достигается хорошее разбиение на два класса, но результаты разбиения на большее число классов оказываются некорректными. Следовательно, с содержательной точки зрения, при минимизации функционала $Q'_1(P)$ производится минимизация разницы между контекстной и прямой близостью объектов. Подобная интерпретация функционала $Q'_1(P)$, предложенного Э. Г. Распини, не является сложной для понимания; детальный же анализ математиче-

ской постановки задачи и условий, предъявляемых к группообразующей функции, является важным с точки зрения установления взаимосвязи между общим видом функционала $Q^i(P)$ и функционалом Э. Г. Распини $Q_1^i(P)$. Уместно также привести замечание И. Д. Манделя о том, что из построения функционала $Q_1^i(P)$ видно, что его трактовка в терминологии теории нечетких множеств «не является принципиальной» [31, с.89].

Для анализа других функционалов качества разбиения достаточно только общего вида функционала (3.12) и ограничений (3.13). Если в формуле (3.12) положить

$$f(p(x_i), v_{ii}) = v_{ii}^2/n, v_{ii} = \mu_{ii}, g(p(x_j), \mu_{ij}) = \mu_{ij}^2/n, \mu_{ij} = v_{ij} \quad (3.28)$$

и

$$\sum_{i=1}^c \mu_{ii} = 1, \sum_{j=1}^n v_{ij} = 1, \quad (3.29)$$

то получится функционал, предложенный в 1985 году М. П. Уиндхемом [186]:

$$Q_2^i(P) = \sum_{i=1}^c \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 v_{ij}^2 d(x_i, x_j), \quad (3.30)$$

где функции принадлежности μ_{ii} и v_{ij} строятся в соответствии со следующими соотношениями:

$$\mu_{ii} = \frac{1 / \sum_{j=1}^n v_{ij}^2 d(x_i, x_j)}{\sum_{k=1}^c \left(1 / \sum_{j=1}^n v_{kj}^2 d(x_i, x_j) \right)}, \quad (3.31)$$

$$v_{ij} = \frac{1 / \sum_{i=1}^c \mu_{ii}^2 d(x_i, x_j)}{\sum_{j=1}^n \left(1 / \sum_{i=1}^c \mu_{ii}^2 d(x_i, x_j) \right)}. \quad (3.32)$$

Как и в прочих критериях качества разбиения, в $Q_2^i(P)$ функция μ_{ii} является функцией принадлежности объекта $x_i \in X$ нечеткому кластеру $A^i \in P$. Вспомогательная же функция принадлежности v_{ij} выражает вес прототипа так, что $v_{ij} = 1$, если некоторый объект $x_j \in X$ в полной мере является прототипом, то есть фактически центром нечеткого кластера $A^i \in P$, и $v_{ij} = 0$ в противном случае. Таким образом, нечеткой является не только степень принадлежности объекта классу,

но и сам представитель класса — каждый элемент классифицируемого множества $X = \{x_1, \dots, x_n\}$ в различной степени может быть, с одной стороны, прототипом того или иного класса, а с другой — просто элементом этого и всех остальных классов. Очевидно, функционал М. П. Уиндхема $Q_2'(P)$ представляет собой наиболее сильное обобщение среди всех оптимизационных методов решения нечеткой модификации задачи кластерного анализа и с содержательной точки зрения, минимизация $Q_2'(P)$ отыскивает оптимальное отклонение нечетких принадлежностей объектов от нечетких центров кластеров. Результатом работы алгоритма, доставляющего минимум функционалу $Q_2'(P)$, являются матрица разбиения $P_{con} = [\mu_{li}]$, где $l = 1, \dots, c, i = 1, \dots, n$, и матрица весов прототипов $K_{con} = [v_{li}]$, где также $l = 1, \dots, c, i = 1, \dots, n$, совместный анализ которых позволяет детально исследовать структуру множества объектов X .

Если же в формуле (3.12) положить

$$f(p(x_i), v_{li}) = v_{li}^2 / n, v_{li} = \mu_{li}, g(p(x_j), \mu_{lj}) = \mu_{lj}^2 / n \quad (3.33)$$

и

$$\sum_{l=1}^c \mu_{li} = 1, \quad (3.34)$$

то получается функционал М. Рубенса [148]

$$Q_3'(P) = \sum_{l=1}^c \sum_{i=1}^n \sum_{j=1}^n \mu_{li}^2 \mu_{lj}^2 d(x_i, x_j), \quad (3.35)$$

минимизация которого отыскивает оптимальное взаимное отклонение элементов в каждом нечетком кластере.

В свою очередь, если положить в формуле (3.12)

$$f(p(x_i), v_{li}) = v_{li} / n, g(p(x_j), \mu_{lj}) = \mu_{lj} / n \quad (3.36)$$

и

$$\sum_{l=1}^c \mu_{li} = 1, \sum_{i=1}^n v_{li} = \text{card}(A^l), \quad (3.37)$$

где $\text{card}(A^l)$ — число элементов в кластере A^l , получается функционал Е. Дидье:

$$Q_4'(P) = \sum_{l=1}^c \sum_{i=1}^n \sum_{j=1}^n v_{li} \mu_{lj} d(x_i, x_j). \quad (3.38)$$

М. Рубенс [148, с.243-244] отмечал, что классификация, получаемая при использовании функционала $Q_4'(P)$, предложенного Е. Дидье, является четкой, что очевидно из второго ограничения (3.37).

Однако можно положить, что если для нескольких точек исследуемой совокупности объектов X выполняется $d(x_i, x_j) = 0, i, j = 1, \dots, n, i \neq j$, то эти точки представляют собой ядро некоторого нечеткого кластера $A^l, l = 1, \dots, c$, так что второе ограничение в условии (3.37) можно будет переписать в виде $\sum_{i=1}^n v_{li} = \text{card}(\text{Core}(A^l)), l = 1, \dots, c, i = 1, \dots, n$. Таким образом, в нечеткой модификации минимизация функционала $Q_4^l(P)$ отыскивает оптимальные отклонения от ядер нечетких кластеров нечетких принадлежностей элементов, не являющихся элементами ядер.

Если же исходные данные представлены матрицей «объект-свойство» $X_{m \times n} = [x_i^t], i = 1, \dots, n, t = 1, \dots, m$, то функционал качества разбиения имеет следующий общий вид [148, с.240]:

$$Q^l(P) = \sum_{i=1}^c \sum_{i=1}^n \sum_{i=1}^n g(p(x_i), \mu_{li}) d(x_i, \tau^l), \quad (3.39)$$

где, как и при рассмотрении функционала $Q^l(P)$, g — некоторый функционал; $p(x_i)$ — априорный вес для каждой точки $x_i, \sum_{i=1}^n p(x_i) = 1$; μ_{li} — функция принадлежности элемента $x_i \in X$ некоторому нечеткому кластеру $A^l \in \{A^1, \dots, A^c\}$; $d(x_i, \tau^l)$ — функция, определяющая расстояние от элемента $x_i, x_i \in X, i = 1, \dots, n$ до некоторого элемента $\tau^l, \tau^l \in X, l = 1, \dots, c$, являющегося центром или, иначе говоря, прототипом нечеткого кластера $A^l, l = 1, \dots, c$, а в качестве ограничений выступают условия

$$\mu_{li} \geq 0, \sum_{i=1}^c \mu_{li} = 1, i = 1, \dots, n, l = 1, \dots, c, \quad (3.40)$$

так что решение задачи классификации заключается в решении оптимизационной задачи

$$Q^l(P) \rightarrow \min \quad (3.41)$$

с соответствующими ограничениями.

Если в формуле (3.39) положить

$$g(p(x_i), \mu_{li}) = \mu_{li}^\gamma / n, 1 \leq \gamma \leq \infty \quad (3.42)$$

и

$$\tau^l = \sum_{i=1}^n \mu_{li}^\gamma \cdot x_i / \sum_{i=1}^n \mu_{li}^\gamma, \quad (3.43)$$

где γ представляет собой показатель нечеткости классификации, то получается функционал Дж. Беждека — Дж. Данна, алгоритм минимизации которого был предложен в работе [83] и обобщен в исследовании [59]:

$$Q_1''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^{\gamma} d(x_i, \tau^l). \quad (3.44)$$

Функционал Дж. Беждека — Дж. Данна представляет собой взвешенный вариант критерия качества разбиения, предложенного в 1965 году М. И. Шлезингером [44], минимизация которого, в свою очередь, отыскивала оптимальные суммарные отклонения координат объектов от центров классов. Поскольку для показателя нечеткости классификации γ выполняется условие $1 \leq \gamma \leq \infty$, алгоритм, оптимизирующий функционал Дж. Беждека — Дж. Данна и именуемый методом нечетких c -средних (fuzzy c -means), представляет собой параметрическое семейство по γ при фиксированном числе кластеров c [37, с.244].

При $\gamma = 1$ полученное разбиение является четким, то есть каждый объект исследуемой совокупности оказывается однозначно принадлежащим какому-либо одному кластеру:

$$\mu_{li} = \begin{cases} 1, & x_i \in A^l \\ 0, & x_i \notin A^l \end{cases}, \forall l = 1, \dots, c, i = 1, \dots, n. \text{ Исследуя поведение функционала } Q_1''(P) \text{ при различных значениях параметра } \gamma, \text{ Дж. Данн [83], [84], [85] предложил } \gamma = 2, \text{ при котором достигается наилучшее разбиение, поскольку при увеличении значения } \gamma \text{ возрастает неопределенность классификации, затрудняющая интерпретацию результатов, что можно выразить следующим образом:}$$

$$\gamma \rightarrow \infty \Rightarrow \mu_{li} \rightarrow \frac{1}{c}, \forall l = 1, \dots, c, \forall i = 1, \dots, n.$$

В качестве функции расстояния в функционале $Q''(P)$, как правило, используется квадрат евклидовой нормы в m -мерном признаковом пространстве $I^m(X)$:

$$d(x_i, \tau^l) = \|x_i - \tau^l\|^2, \quad (3.45)$$

так что формула (3.44) может быть переписана, соответственно, в следующем виде:

$$Q_1''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^{\gamma} \|x_i\|^2 \quad (3.46)$$

Следует, однако, указать, что использование евклидова расстояния в функционале $Q_1''(P)$ оправдано в тех случаях, когда нечеткие кластеры имеют примерно одинаковый размер и форму гиперсферы.

При исследовании функционала Дж. Беждека — Дж. Данна в качестве функции $d(x_i, \tau^l)$ Д. Е. Густафсоном и В. С. Кесселем [92] рассматривался квадрат расстояния Махаланобиса:

$$d(x_i, \tau^l) = \left(\sqrt{(x_i - \tau^l)^T C_l^{-1} (x_i - \tau^l)} \right)^2, \quad (3.47)$$

так что

$$|C_l| = 1, l = 1, \dots, c, \quad (3.48)$$

$$C_l = |U_l|^{-1/m} U_l^{-1}, \quad (3.49)$$

$$U_l = \left(\sum_{i=1}^n \mu_{ii}^{\gamma} (x_i - \tau^l)(x_i - \tau^l)^T \right) / \left(\sum_{i=1}^n \mu_{ii}^{\gamma} \right), \quad (3.50)$$

где U_l — ковариационная матрица кластера A^l , то есть некоторая симметричная неотрицательно-определенная матрица, и T — символ транспонирования. При таком подходе допускается пересечение нечетких кластеров, однако кластеры должны иметь гиперэллиптическую форму. Уместно указать, что при выполнении условия $C_l = I, \forall l$ расстояние $d(x_i, \tau^l)$, определяемое выражением (3.47), будет представлять собой квадрат евклидовой нормы (3.45).

Кроме того, П. Гроененом и К. Яюгой при исследовании функционала $Q_1''(P)$ в качестве функции $d(x_i, \tau^l)$ было предложено использование расстояния Миньковского [106, с.415]. Более того, К. Яюгой в работе [105] рассматривается проблема применения L_1 -метрики в методе нечетких c -средних, а С. Мийамото и Ю. Агуста в работе [129] предлагают быстрый алгоритм вычисления центров нечетких кластеров в методе нечетких c -средних в L_1 -пространстве; вычислительные эксперименты проводились в сравнении с методом нечетких c -средних, использующим классическую евклидову метрику для 10 000 объектов, и показали высокую эффективность процедуры. Дж. Беждек, К. Корэй, Р. Гундерсон и Дж. Уотсон в работе [60] также предлагают несколько модификаций метода нечетких c -средних. Особо следует отметить результаты, описанные С. А. Айвазяном, В. М. Бухштабером, И. С. Енюковым и Л. Д. Мешалкиным в исследовании [37, с.291-300].

С содержательной точки зрения интерпретация функционала Дж. Беждека — Дж. Данна оказывается весьма простой: минимизация

$Q_1''(P)$ отыскивает оптимальные нечеткие отклонения объектов $x_i \in X, i=1, \dots, n$ от прототипов $\tau^l, l=1, \dots, c$ нечетких кластеров $A^l, l=1, \dots, c$, или, как отмечал профессор Л. А. Заде, функционал $Q_1''(P)$ «выполняет роль оценки взвешенной дисперсии точек из X относительно оптимального расположения центров $\tau^1, \dots, \tau^c$ » [25, с.231]. Следует также отметить то обстоятельство, что функционал Дж. Беждека — Дж. Данна, как отмечает И. Д. Мандель [31, с.92], «представляет собой, видимо, наиболее распространенный и изученный вариант экстремальной постановки задачи кластер-анализа в терминах размытых множеств».

Пусть X_L — множество помеченных объектов, $X_L \subset X$, элементы которого представлены булевыми векторами $s = (s_1, s_2, \dots, s_n)^T$, где T — символ транспонирования и

$$s_i = \begin{cases} 1, x_i \in X_L \\ 0, x_i \notin X_L \end{cases}, \quad (3.51)$$

а $Y_{\text{сxn}} = [y_{li}], l=1, \dots, c, i=1, \dots, n$ — матрица разбиения, составляемая исследователем в соответствии со следующим правилом: если $x_i \in X_L$, то y_{li} задается с соблюдением условия $\sum_{l=1}^c y_{li} = 1$, где y_{li} — степень принадлежности помеченного объекта $x_i, x_i \in X_L$ классу $A^l, l=1, \dots, c$; в противном случае, то есть если $x_i \notin X_L$, соответствующий столбец в матрице $Y_{\text{сxn}}$ оказывается неуместным и исследователь может его пропустить при рассмотрении [144, с.142]. Тогда, полагая в формуле (3.39)

$$g(p(x_i), \mu_{li}) = \mu_{li}^2 + (\mu_{li} - s_{li} y_{li})/n, \quad (3.52)$$

и

$$\tau^l = \sum_{i=1}^n \mu_{li}^2 \cdot x_i / \sum_{i=1}^n \mu_{li}^2, \quad (3.53)$$

получаем функционал В. Педрича [140]

$$Q_2''(P) = \sum_{l=1}^c \sum_{i=1}^n (\mu_{li}^2 + (\mu_{li} - s_{li} y_{li})^2) d(x_i, \tau^l), \quad (3.54)$$

который может быть переписан в следующем виде:

$$Q_2''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^2 d(x_i, \tau^l) + \sum_{l=1}^c \sum_{i=1}^n (\mu_{li} - s_{li} y_{li})^2 d(x_i, \tau^l). \quad (3.55)$$

Если в качестве функции расстояния в функционале $Q_2''(P)$ использовать квадрат евклидовой нормы (3.45), то формула (3.54) может быть записана в следующем виде:

$$Q_2''(P) = \sum_{l=1}^c \sum_{i=1}^n (\mu_{li}^2 + (\mu_{li} - s_{li} y_{li})^2) \|x_i - \tau^l\|^2, \quad (3.56)$$

а формула (3.55), соответственно, в виде

$$Q_2''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^2 \|x_i - \tau^l\|^2 + \sum_{l=1}^c \sum_{i=1}^n (\mu_{li} - s_{li} y_{li})^2 \|x_i - \tau^l\|^2. \quad (3.57)$$

В. Педрич исследовал различные модификации критерия (3.54) [140], первая из которых связана со взвешиванием обоих слагаемых в выражении (3.55), что приводит к следующей форме критерия $Q_2''(P)$:

$$Q_2''(P) = a_2 \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^2 d(x_i, \tau^l) + a_2 \sum_{l=1}^c \sum_{i=1}^n (\mu_{li} - s_{li} y_{li})^2 d(x_i, \tau^l), \quad (3.58)$$

где $a_1 = 1/(n - n_L)$, $a_2 = 1/n_L$ и $n_L = \text{card}(X_L)$, причем в качестве функции $d(x_i, \tau^l)$ в выражении (3.58) использовался квадрат евклидовой нормы (3.45).

Помеченные образы могут использоваться для изменения функции $d(x_i, \tau^l)$ различным образом. Очевидно, что помеченные образы частично определяют внутреннюю структуру исследуемой совокупности $X = \{x_1, \dots, x_n\}$ и могут быть использованы для построения нормы в соответствующем m -мерном признаковом пространстве $I^m(X)$, для чего вводится матрица с помеченными объектами C_l :

$$C_l = \frac{\sum_{i=1}^n y_{li}^2 s_i (x_i - \tau^l)(x_i - \tau^l)^T}{\sum_{i=1}^n y_{li}^2 s_i}, \quad (3.59)$$

где T — символ транспонирования, а нечеткие средние τ^l вычисляются в соответствии с формулой

$$\tau^l = \sum_{i=1}^n y_{li}^2 s_i x_i \sum_{i=1}^n y_{li}^2 s_i, l = 1, \dots, c, \quad (3.60)$$

что подразумевает введение в качестве функции $d(x_i, \tau^l)$ в выражение (3.54) квадрата расстояния Махаланобиса в виде

$$d(x_i, \tau^l) = \left(\sqrt{(x_i - \tau^l)^T C_l^{-1} (x_i - \tau^l)} \right)^2 = \|x_i - \tau^l\|_{C_l^{-1}}^2, \quad (3.61)$$

что, в свою очередь, дает вторую модификацию критерия $Q_2''(P)$.

Третьей модификацией критерия $Q_2''(P)$ является функционал (3.58) с введением в качестве функции $d(x_i, \tau')$ квадрата расстояния Махаланобиса (3.61).

С содержательной точки зрения множество помеченных объектов X_L представляет собой частично обучающую выборку, элементы которой являются своего рода «эталоны» для классификации. Смысл же собственно функционала $Q_2''(P)$ заключается в следующем: минимизация первого слагаемого в формуле (3.55), полностью совпадающего с $Q_1''(P)$ при $\gamma = 2$ в формуле (3.44), минимизирует нечеткие суммы квадратов расстояний от объектов до центров нечетких кластеров, а второе слагаемое в (3.55) является взвешенной по квадратам расстояний суммой отклонений расчетных значений функции принадлежности объектов нечетким кластерам от заданных априорно.

Особо следует отметить функционал качества разбиения, предложенный В. Райтом [189]:

$$Q_3''(P) = 1 - \frac{\sum_{l=1}^c \sum_{i=1}^n \mu_{li} d^2(x_i, \tau^l)}{\sum_{l=1}^c \sum_{i=1}^n \mu_{li} d^2(x_i, \bar{x})}, \quad (3.62)$$

где $d^2(x_i, \bar{x})$ — квадрат расстояния от точки x_i до общего центра исследуемой совокупности, который может быть определен в соответствии с формулой

$$d^2(x_i, \bar{x}) = \sum_{i=1}^m \left(x_i' - \frac{1}{n} \sum_{j=1}^n x_j' \right)^2, \quad (3.63)$$

а $d^2(x_i, \tau^l)$ — квадрат расстояния от точки x_i до центра τ^l нечеткого кластера A^l , определяемый из соотношения

$$d^2(x_i, \tau^l) = \sum_{i=1}^m \left(x_i' - \frac{\sum_{j=1}^n \mu_{lj} x_j'}{\sum_{j=1}^n \mu_{lj}} \right)^2. \quad (3.64)$$

Нахождение оптимальной классификации в данном случае определяется как решение следующей оптимизационной задачи:

$$Q_3''(P) \rightarrow \max, \quad (3.65)$$

где в качестве ограничений выступают условия (3.40).

Выражение (3.62) представляет собой, по сути, взвешенный аналог корреляционного отношения и сравнивает внутриклассовые

расстояния с общими расстояниями. Вместе с тем, соотношения (3.63) и (3.64) для построения центрального элемента классифицируемого множества и центров классов применимы только в случае, когда все признаки x_i^l описания элементов исследуемой совокупности измерены на интервальных шкалах, а пространство признаков является евклидовым [22, с.58].

Следует указать, что функционал $Q_3''(P)$ прямо не связан с общим видом функционала $Q''(P)$, так же, как и различаются между собой задачи (3.41) и (3.65). Общим для функционалов $Q''(P)$ и $Q_3''(P)$ является только вид матрицы исходных данных. Вместе с тем, функционал $Q_3''(P)$ показал хорошие свойства относительно 12 аксиом, выдвинутых В. Райтом [189], которым, по его мнению, должен удовлетворять функционал $Q''(P)$; при следовании этим положениям, по свидетельству В. Райта, в процессе выбора функционала качества форма $Q''(P)$ оказывается наиболее адекватной. Следует, однако, указать, что требования, выдвинутые В. Райтом, относятся к критериям качества разбиения вообще, а не только нечеткой классификации.

В. Райт рассматривает критерий качества разбиения как заданную на множестве упорядоченных пар $(x_i^l, \mu_{il}), i = 1, \dots, n, l = 1, \dots, m, l = 1, \dots, c$ функцию, где, как и ранее, n — число элементов классифицируемого множества $X = \{x_1, \dots, x_n\}$; m — число признаков, или, иными словами, размерность пространства $I^m(X)$; c — число нечетких кластеров — элементов множества $P = \{A^1, \dots, A^c\}$; x_i^l — элемент матрицы «объект-свойство» $X_{m \times n} = [x_i^l], i = 1, \dots, n, l = 1, \dots, m$, выражающий значение l -го признака для i -го объекта; μ_{il} — элемент матрицы разбиения $P_{c \times n} = [\mu_{il}], l = 1, \dots, c, i = 1, \dots, n$, представляющий степень принадлежности l -му нечеткому кластеру i -го элемента исследуемого множества.

Целесообразно указать некоторые ограничения, накладываемые на матрицы $P_{c \times n}$ и $X_{m \times n}$. В первую очередь необходимо отметить следующее обстоятельство: если в i -м столбце матрицы разбиения $P_{c \times n}$ один или несколько элементов μ_{il} принимают значение, равное нулю, то ни одна из l -х строк не может содержать все нули, так как в этом случае подмножество A^l оказывается пустым, что противоречит определению понятия кластера. Во вторую очередь следует указать, что

в матрице разбиения $P_{c \times n}$ не может быть нулевых столбцов и число кластеров должно отвечать условию $2 \leq c \leq n$, поскольку в случае $c > n$ образуются пустые кластеры, а в случае $c = 1$ получается единственный кластер, так что невозможно оценить его сходство или отличие от других кластеров, что, в свою очередь, противоречит принципам построения любой классификации.

Таким образом, функционал качества $Q''(P)$ должен удовлетворять следующим основным, по мнению В. О. Рукавишникова [22, с.50-51], из двенадцати сформулированных В. Райтом, требованиям:

- 1) $Q''(P)$ должен быть инвариантен к изменению порядка строк в матрице $P_{c \times n}$;
- 2) $Q''(P)$ должен быть инвариантен к изменению порядка строк и столбцов в матрице $X_{m \times n}$;
- 3) $Q''(P)$ должен быть инвариантен по отношению к объединению элементов в соответствии со следующим положением: любые два элемента классифицируемой совокупности $X = \{x_1, \dots, x_n\}$, которые имеют одни и те же координаты в признаковом пространстве $I^m(X)$ и различные степени принадлежности к нечетким кластерам, эквивалентны одному элементу с теми же координатами и степенями принадлежности к соответствующим нечетким кластерам, равным половинам сумм степеней принадлежности к некоторому данному нечеткому кластеру объединяемых элементов.
- 4) $Q''(P)$ не должен зависеть от объединения идентичных по положению в пространстве $I^m(X)$ и по степеням принадлежности μ_{ii} к нечетким кластерам нечеткого разбиения $P = \{A^1, \dots, A^c\}$ элементов.

Касательно второго требования следует указать, что по отношению к столбцам матрицы $X_{m \times n}$ В. Райтом выдвигается более жесткое требование, в соответствии с которым функционал $Q''(P)$ не должен зависеть от преобразований координатных осей — признаков, к примеру, вращения или изменения шкал.

Безусловно, рассмотренные выше критерии качества нечеткого разбиения не исчерпывают всего многообразия предложенных функционалов. Другие критерии рассматриваются в работах [46], [59], [64], [85], [101], [108], [167], [185] и, в основном, представляют собой раз-

личные версии рассмотренных выше функционалов. В частности, в работе Р. Н. Даве и С. Сена [73] рассматривается функционал

$$Q'_5(P) = \frac{\sum_{i=1}^c \sum_{j=1}^n \sum_{l=1}^n \mu_{il}^{\gamma} \mu_{lj}^{\gamma} d(x_i, x_j)}{2 \sum_{k=1}^n \mu_{ik}^{\gamma}}, \quad (3.66)$$

отличающийся от функционала, предложенного Л. Кофманом и П. Дж. Рауссеу [108] тем, что в критерии Л. Кофмана — П. Дж. Рауссеу значение показателя нечеткости классификации γ полагается равным двум, тогда как для критерия Р. Н. Даве и С. Сена $Q'_5(P)$ значение показателя γ полагается более общим, так что выполняется условие $1 < \gamma < \infty$.

Представляется целесообразным кратко рассмотреть такую весьма актуальную проблему, как робастность нечетких методов кластерного анализа. Термин «робастность», имеющий английское происхождение, означает устойчивость кластер-процедур к нарушениям модельных предположений о классифицируемых наблюдениях. Среди наиболее распространенных искажений следует отметить аномальные наблюдения, именуемые также «выбросами» или «шумом», пропуски значений признаков, зависимость наблюдений. Робастные методы получили весьма широкое распространение в вероятностных методах классификации. Среди зарубежных исследователей проблемы робастности в статистике следует указать таких авторов, как Дж. В. Тьюки, П. Хьюбер, П. Дж. Рауссеу, Ф. Р. Хампель, К. А. Дучинскас, Ш. Ю. Раудис, а также отметить труды известных отечественных специалистов: С. А. Айвазяна [37], Ю. С. Харина [110] и Е. Е. Жука [23].

Исследование проблемы робастности получило развитие также и в рамках нечеткого подхода к решению задачи распознавания образов с самообучением, несмотря на его новизну в сравнении с геометрическим и вероятностно-статистическим подходами к решению этой задачи. В силу ограниченности изложения рассмотрение проблемы робастности нечетких кластер-процедур будет проводиться на основе результатов, представленных Р. Н. Даве и С. Сеном в работе [73].

«Выбросы» могут рассматриваться как отдельный кластер A^* с центром τ^* , расстояние до которого от всех объектов исследуемой совокупности неизменно, так что $d(x_i, \tau^*) = \delta, \forall x_i \in X$. Тогда функция принадлежности объектов x_i исследуемой совокупности X классу A^* может определяться в соответствии с формулой

$$\mu_{ai} = 1 - \sum_{l=1}^c \mu_{li}, \quad (3.67)$$

так что c будет представлять собой число «хороших» нечетких кластеров, то есть кластеров, не содержащих аномальных наблюдений. Использование соотношения (3.67) приводит к следующему ослаблению условия нечеткого разбиения в смысле Э. Г. Распини:

$$\sum_{l=1}^c \mu_{li} \leq 1, l=1, \dots, c, i=1, \dots, n. \quad (3.68)$$

Таким образом, точки класса A^* будут иметь малые значения принадлежности «хорошим» кластерам, а кластер A^* представляет собой, в сущности, класс «джокер» [37, с.290-291]. В свою очередь, функционал Дж. Беждека — Дж. Данна (3.44) может быть переписан в виде

$$Q_{1R}''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^\gamma d(x_i, \tau^l) + \sum_{i=1}^n \delta^2 \left(1 - \sum_{a=1}^c \mu_{ai} \right)^\gamma, \quad (3.69)$$

где символ R в обозначении функционала обозначает его робастность. При нахождении экстремума функционала $Q_{1R}''(P)$ значения принадлежности $\mu_{li}, l=1, \dots, c, i=1, \dots, n$ отыскиваются в соответствии с формулой

$$\mu_{li} = \frac{\left(\frac{1}{d(x_i, \tau^l)} \right)^{1/(\gamma-1)}}{\sum_{a=1}^c \left(\frac{1}{d(x_i, \tau^a)} \right)^{1/(\gamma-1)} + \left(\frac{1}{\delta^2} \right)^{1/(\gamma-1)}}, \quad (3.70)$$

отличающейся от аналогичной в процедуре минимизации критерия $Q_1''(P)$ наличием второго слагаемого в делителе в правой части выражения (3.70). Как отмечают Р. Н. Даве и С. Сен, «выбор δ является сложной проблемой» [73, с.716], так как, в принципе, требуется знание о размере каждого кластера. Однако во многих практических задачах расстояние до «выбросов» является постоянным, так что значение δ может быть определено исходя из выражения

$$\delta^2 = \varpi \left[\frac{\sum_{l=1}^c \sum_{j=1}^n (d(x_j, \tau^l))^2}{cn} \right], \quad (3.71)$$

где множитель ϖ выбирается в зависимости от типа данных [73, с.716].

Кластеризация данных с «выбросами» хорошо применима в ситуациях, когда исходные данные представлены в форме матрицы «объект-свойство», где расстояние δ имеет вполне ясный смысл. Однако в случае, когда исходные данные представлены в форме матрицы «объект-объект», существуют некоторые очевидные трудности выделения класса «выбросов» A^* . При рассмотрении робастности критериев вида $Q^l(P)$ в работе [73] предлагается следующая модификация функционала М. Рубенса:

$$Q_3^l(P) = \sum_{i=1}^{c+1} \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 \mu_{ij}^2 d(x_i, x_j), \quad (3.72)$$

где число c отыскиваемых кластеров увеличивается на единицу. Этот дополнительный класс и будет классом «выбросов» A^* , однако из выражения (3.72) не ясно, каким образом класс A^* выделяется из остальных кластеров. Если выражение (3.72) переписать в виде

$$Q_3^l(P) = \sum_{l=1}^c \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 \mu_{ij}^2 (d(x_i, x_j))_l + \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 \mu_{ij}^2 (d(x_i, x_j))_*, \quad (3.73)$$

то первое слагаемое в правой части совпадает с функционалом М. Рубенса (3.35), а второе представляет собой своего рода «расширение» для классификации «выбросов». Индекс l в расстоянии $(d(x_i, x_j))_l$ обозначает «существенность» различия объектов x_i и x_j , так, как это определяется кластером A^l . В обычном случае различие между элементами исследуемой совокупности не зависит от кластера, так что $(d(x_i, x_j))_l = d(x_i, x_j)$ для всех $A^l, l=1, \dots, c$, однако в случае, когда вводится класс «выбросов», его следует выделять как специфический кластер A^* и налагать на него приоритет определения существенно-сти различия между объектами, так что $(d(x_i, x_j))_* = \delta$, и, соответственно, класс «выбросов» A^* в случае представления исходных данных в виде матрицы «объект-объект» может быть определен как кластер, расстояние между объектами которого оценивается так же, как и до объектов других кластеров. Иными словами, кластер A^* определяется так, что $(d(x_i, x_j))_* = \delta$ для всех x_i и x_j , и расстояние δ представляет собой константу. Следовательно, выражение (3.73) можно переписать так, что робастная модификация функционала М. Рубенса окончательно примет следующий вид:

$$Q_{3R}^l(P) = \sum_{l=1}^c \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 \mu_{ij}^2 d(x_i, x_j) + \sum_{i=1}^n \sum_{j=1}^n \mu_{ii}^2 \mu_{ij}^2 \delta. \quad (3.74)$$

Таким образом, становится ясным смысл расстояния δ с содержательной точки зрения: оптимальное разбиение исследуемой совокупности таково, что различие между объектами в «хороших» классах сравнительно невелико, а в классе «выбросов» все коэффициенты попарных различий между объектами принимают одно и то же значение δ ; таким образом, в каждом «хорошем» кластере ни одна пара объектов не имеет степени различия, большей чем δ , поскольку, если какой-либо объект исследуемой совокупности отличается от каждого объекта в каждом «хорошем» кластере более чем на δ , он должен быть классифицирован как «выброс», и напротив, если этот объект отличается хотя бы от одного объекта, принадлежащего «хорошему» кластеру, менее чем на δ , рассматриваемый объект должен быть приписан этому «хорошему» кластеру, а не классу «выбросов». В качестве иллюстративного примера Р. Н. Даве и С. Сен [73, с.717] рассматривают совокупность из семи объектов $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$, где первые три объекта представляют собой яблоки трех различных сортов, вторые три объекта представляют собой цитрусовые трех сортов, а седьмой объект — сливу. Матрица коэффициентов различия объектов исследуемой совокупности X представлена таблицей 3.1.

Таблица 3.1

Матрица попарных расстояний между объектами исследуемой совокупности

I	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0.00						
x_2	0.10	0.00					
x_3	0.15	0.20	0.00				
x_4	0.45	0.60	0.55	0.00			
x_5	0.50	0.45	0.50	0.15	0.00		
x_6	0.45	0.50	0.45	0.20	0.10	0.00	
x_7	0.40	0.45	0.60	0.45	0.50	0.55	0.00

В данном простом примере, для иллюстрации понятия расстояния δ отличия «выбросов», значение δ можно положить равным 0.30, так что хорошо выделяются классы $\{x_1, x_2, x_3\}$ и $\{x_4, x_5, x_6\}$, расстояния внутри которых варьируются в пределах от 0.10 до 0.20, тогда как степень отличия объекта x_7 от остальных объектов превышает $\delta = 0.30$, что позволяет классифицировать объект x_7 как «выброс». В реаль-

ных задачах значение расстояния δ можно вычислить по формуле, аналогичной соотношению (3.71).

Исходя из вышеизложенного, робастная модификация функционала $Q'_5(P)$ Р. Н. Даве — С. Сена может быть записана в следующем виде [73, с.718]:

$$Q'_{SR}(P) = \sum_{l=1}^c \frac{\sum_{j=1}^n \sum_{i=1}^n \mu_{ij}^{\gamma} \mu_{ij}^{\gamma} d(x_i, x_j)}{2 \sum_{k=1}^n \mu_{lk}^{\gamma}} + \frac{\sum_{j=1}^n \sum_{i=1}^n \mu_{ij}^{\gamma} \mu_{ij}^{\gamma} \delta}{2 \sum_{k=1}^n \mu_{*k}^{\gamma}}. \quad (3.75)$$

Р. Н. Даве и С. Сен отмечают [73, с.718], что аналогичным путем, то есть введением класса «выбросов» и соответствующего расстояния δ , получается робастная модификация функционала $Q'_2(P)$ М. П. Уиндхема.

В завершение рассмотрения нечетких оптимизационных методов автоматической классификации необходимо отметить, что во всех рассмотренных выше критериях качества нечеткого разбиения учитывается число кластеров, так что число кластеров является параметром, который необходимо задавать при работе того или иного оптимизационного алгоритма. Данное число задается исследователем, исходя из сущности решаемой задачи. Вместе с тем, необходимо, чтобы число классов отвечало «естественному» расслоению объектов по классам, так что в ряде случаев отыскание «естественного» числа классов представляет собой довольно серьезную проблему, что более детально будет рассматриваться ниже. Предварительно следует лишь указать, что для решения проблемы отыскания «приемлемого» числа кластеров используются два подхода. Первый подход состоит в проведении ряда экспериментов для различного числа классов в искомом нечетком разбиении с последующим выбором нечеткого разбиения, число классов в котором отвечает формальным или содержательным представлениям исследователя о «естественности» числа классов. Второй подход состоит в задании заведомо большого числа классов ввиду того, что в процессе классификации наиболее близкие классы объединяются. Другим обстоятельством, на которое необходимо обратить внимание, является то, что в нечеткой кластеризации принадлежность точек кластерам уменьшается с возрастанием расстояния до точки — центра кластера. Вместе с тем, во многих практических задачах точки, вплотную прилегающие к прототипу нечеткого кластера, могут рассматриваться как полностью принадлежащие нечеткому множеству, представленному соответствующим кластером. Данное обстоятельство

во диктует необходимость расширения прототипа кластера на определенное расстояние от центра кластера, так чтобы точки, попавшие в область расширения прототипа, считались бы принадлежащими соответствующему кластеру со степенью принадлежности 1.0. Все функционалы, рассмотренные выше, такой возможности исследователю не предоставляют.

Решение обозначенной проблемы «расширения» прототипа кластера и отыскания «естественного» числа классов в нечетком разбиении было предложено У. Кеймаком и М. Сетнесом в работе [109] путем введения понятия объемного прототипа и адаптивного объединения ближайших нечетких кластеров. Объемный прототип $\tilde{\tau}^l \in I^m(X)$ некоторого нечеткого кластера $A^l \in P$ может быть определен как n -мерное выпуклое и компактное подпространство и, таким образом, расширяет прототип кластера от единичной точки до области в пространстве $I^m(X)$ [109, с.706]. Следует указать на то, что подобное определение предполагает произвольную форму и размер объемного прототипа. Таким образом, объемный прототип образуется путем включения в область $\tilde{\tau}^l \in I^m(X)$ всех точек в радиусе R_l от центра τ^l нечеткого кластера $A^l \in P$. Точки $x_i \in X$, удовлетворяющие неравенству $d(x_i, \tau^l) \leq R_l$, будут являться элементами объемного прототипа $\tilde{\tau}^l$, так что, по определению, их степень принадлежности кластеру A^l будет максимальной. Размер объемного прототипа $\tilde{\tau}^l$ будет определяться радиусом R_l , который, в свою очередь, может задаваться исследователем, так что все объемные прототипы $\tilde{\tau}^l, l=1, \dots, c$ будут иметь фиксированный размер либо определяться для каждого кластера $A^l, l=1, \dots, c$ в процессе классификации.

Естественным путем к определению радиуса R_l некоторого кластера $A^l \in P$ является его соотнесение с ковариационной матрицей U_l (3.50), детерминант $|U_l|$ которой определяет объем кластера. Поскольку матрица U_l является неотрицательно-определенной и симметричной, она может быть представлена в виде $U_l = G_l \Lambda_l G_l^T$, где T — символ транспонирования, а матрица Λ_l является диагональной с ненулевыми элементами $\lambda_{l1}, \dots, \lambda_{lm}$, так что объемные прототипы расширяют расстояния $\sqrt{\lambda_{lt}}, t=1, \dots, m$ вдоль каждого вектора g_{lt} .

Кластеризация с объемными прототипами подразумевает минимизацию функционала

$$Q_{1E}''(P) = \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^{\gamma} (d(x_i, \tau^l) - R_l^2), \quad (3.76)$$

представляющего собой модификацию критерия (3.44), где символ E в обозначении $Q_{1E}''(P)$ функционала обозначает «расширенный» (extended), $d(x_i, \tau^l)$ — функция расстояния, в качестве которой может быть использован, к примеру, квадрат расстояния Махаланобиса (3.47), а R_l — постоянная, определяющая размер объемного прототипа нечеткого кластера $A^l \in P$. В качестве ограничений на целевую функцию (3.76) выступает традиционное условие разбиения в смысле Э. Г. Распини. Таким образом, центры $\tau^l, l=1, \dots, c$ нечетких кластеров $A^l, l=1, \dots, c$ отыскиваются по формуле

$$\tau^l = \sum_{i=1}^n \mu_{li}^{\gamma} \cdot x_i / \sum_{i=1}^n \mu_{li}^{\gamma}, l=1, \dots, c, \quad (3.77)$$

а значения принадлежностей, соответственно, по формуле

$$\mu_{li} = \frac{1}{\sum_{a=1}^c \left(\frac{d(x_i, \tau^l) \left(1 - \frac{R_l^2}{d(x_i, \tau^l)} \right)}{d(x_i, \tau^a) \left(1 - \frac{R_a^2}{d(x_i, \tau^a)} \right)} \right)^{1/(\gamma-1)}}, l=1, \dots, c, i=1, \dots, n, \quad (3.78)$$

которая, в свою очередь, может быть переписана в виде

$$\mu_{li} = \frac{1}{\sum_{a=1}^c \left(\frac{d(x_i, \tau^l) \zeta_{li}}{d(x_i, \tau^a) \zeta_{ai}} \right)^{1/(\gamma-1)}}, l=1, \dots, c, i=1, \dots, n, \quad (3.79)$$

где взвешивающая компонента ζ_{li} определяется выражением

$$\zeta_{li} = \max \left(0, 1 - \frac{R_l^2}{d(x_i, \tau^l)} \right). \quad (3.80)$$

При минимизации функционала $Q_{1E}''(P)$ центры $\tau^l, l=1, \dots, c$ кластеров $A^l, l=1, \dots, c$ «расширяются» до объемных прототипов $\tilde{\tau}^l, l=1, \dots, c$, так что точки, попадающие в область объемного прототипа, будут иметь степень принадлежности соответствующему кластеру, равную единице, и, соответственно, нулю — другим кластерам. Поскольку метод, предложенный У. Кеймаком и М. Сетнесом [109], с

целью нахождения «естественного» числа кластеров $A^l, l=1, \dots, c$ в отыскиваемом нечетком разбиении P предполагает объединение ближайших кластеров, радиус кластера умножается на величину $\chi^{(b)}/c^{(b)}$, где $c^{(b)}$ представляет собой число кластеров в разбиении $P_{(b)}$ на b -м шаге вычислительного процесса, а алгоритм начинает работу при $\chi^{(0)} := 1$, так что при объединении кластеров увеличение значения $\chi^{(b)}, \chi^{(b)} \leq c^{(b)}$ делает возможным увеличение размеров объемных прототипов.

Объединение ближайших кластеров производится в соответствии с мерой сходства, определяемой выражением

$$S_{lf} = \frac{\sum_{i=1}^n \min(\mu_{li}, \mu_{fi})}{\min\left(\sum_{i=1}^n \mu_{li}, \sum_{i=1}^n \mu_{fi}\right)}. \quad (3.81)$$

Мера, определяемая выражением (3.81), принимает во внимание подобие точек, как принадлежащих объемным прототипам нечетких кластеров, так и лежащих вне их пределов.

Порог $\alpha, \alpha \in [0,1]$, при котором происходит объединение ближайших кластеров, зависит от различных характеристик исходных данных, таких, к примеру, как делимость групп, плотность кластеров и их размер, а также параметров классификации, в первую очередь от показателя нечеткости γ . В общем, в случае обращения к расширенным методам нечеткой классификации порог объединения α оказывается параметром, определяемым пользователем. Степень «близости» некоторых двух кластеров $A^l, A^f, l, f=1, \dots, c, l \neq f$ зависит также и от других кластеров в нечетком разбиении, что определяется таким ограничением, как условие разбиения в смысле Э. Г. Распини. Однако выбор порога исследователем при решении конкретной задачи может вызывать затруднения. Для решения подобного рода проблемы У. Кеймаком и М. Сетнесом [109] предлагается использовать адаптивный порог $\alpha^{(b)}$, зависящий от числа кластеров $c^{(b)}$ в нечетком разбиении на $P_{(b)}$ на b -м шаге вычислительного процесса, который определяется выражением

$$\alpha^{(b)} = \frac{1}{c^{(b)} - 1}, \quad (3.82)$$

так что нечеткие кластеры A^l и A^f объединяются в случае, когда изменение значений меры сходства кластеров, определяемой выражением

ем (3.81), не превышает некоторого порогового значения $\varepsilon_1 > 0$ и значение меры сходства на b -м шаге процесса превышает значение $\alpha^{(b)}$.

Для вычисления радиуса R_l , определяющего размер объемного прототипа, целесообразно представить ковариационную матрицу U_l в виде $U_l = G_l \Lambda_l G_l^T$, как это было предложено выше, так что с учетом условий (3.48) и (3.49)

$$U_l = |U_l|^{1/m} \mathbf{I}, \quad (3.83)$$

откуда следует, что радиус R_l может быть определен в соответствии с формулой

$$R_l = \sqrt{|U_l|^{1/m}} = \sqrt{\prod_{i=1}^m \lambda_{li}^{1/m}}, \quad (3.84)$$

где λ_{li} — элементы матрицы Λ_l .

3.2.2. Описание алгоритмов

Во всех исследованиях, как правило, приводятся не только функционалы качества разбиения, но и алгоритмы поиска экстремума рассматриваемого функционала.

3.2.2.1. Алгоритм Распини [150] минимизирует критерий $Q_l^i(P)$ в виде (3.27), так что находится решение $P^* = \arg \min_P \left\{ \begin{array}{l} Q_l^i(P) : P = (A^1, \dots, A^c), A^i = (\mu_{i1}, \dots, \mu_{in}), 0 \leq \mu_{il} \leq 1, \\ \sum_{i=1}^c \mu_{il} = 1, \sum_{i=1}^n \mu_{li} > 0, i = 1, \dots, n, l = 1, \dots, c \end{array} \right\}.$

Как отмечали И. И. Елисеева и В. О. Рукавишников касательно функционала Э. Г. Распини, «поиск экстремального значения критерия качества целесообразно проводить с использованием градиентных методов, обеспечивающих быструю сходимость алгоритмов» [22, с.59]. Дальнейшее изложение алгоритма, минимизирующего функционал $Q_l^i(P)$, основано на описании процедуры, представленном в работе [22].

Пусть $V_m = \max V(r(x_i), r(x_j))$, где $V(r(x_i), r(x_j))$ вычисляется по формуле (3.25), а функция $F(d(x_i, x_j))$ вычисляется в соответствии

с соотношением $F(d(x_i, x_j)) = \begin{cases} V_m \left(\frac{d(x_i, x_j)}{\hat{d}} \right)^2, & d(x_i, x_j) \leq \hat{d} \\ V_m, & d(x_i, x_j) > \hat{d} \end{cases}$. Кроме то-

го, для удобства преобразований можно ввести также следующие обозначения:

$$V_{(b)} = V_{(b)}(r(x_i), r(x_j)), \quad (3.85)$$

выражающее значение $V(r(x_i), r(x_j))$ на b -м шаге вычислительного процесса

$$\sigma_b = \frac{1}{V_{(b)m}}, \quad (3.86)$$

где символом $V_{(b)m}$ обозначается $V_m = \max V(r(x_i), r(x_j))$ на b -м шаге вычислительного процесса а также

$$\frac{d(x_i, x_j)}{\hat{d}_{(b)}} = d_{(b)}(x_i, x_j), \quad (3.87)$$

так что на b -м шаге вычислительного процесса задача будет заключаться в нахождении

$$Q_{1(b)}^l(P) = \sum_{i=1}^n \sum_{j=1}^n (\sigma_b V_{(b)} - d_{(b)}^2(x_i, x_j))^2 \rightarrow \min. \quad (3.88)$$

Следует отметить, что $Q_1^l(P)$ представляет собой дифференцируемую функцию, частные производные которой находятся по следующим формулам:

$$\beta_l(x_i) = -\frac{\partial Q_1^l(P)}{\partial \mu_{li}}, l=1, \dots, c, i=1, \dots, n; \quad (3.89)$$

$$\beta_{\max}(x_i) = \max_{(l)} \beta_l(x_i); \quad (3.90)$$

$$\beta_{\min}(x_i) = \min_{(l)} \beta_l(x_i); \quad (3.91)$$

и, поскольку $Q_1^l(P)$ — полином четвертой степени, то его частная производная в точке x_i на b -м шаге вычислительного процесса будет определяться следующим соотношением:

$$\beta_{(b)l}(x_i) = \sum_{j=1}^n \sigma_b p(x_j) (\sigma_b V_{(b)} - d_{(b)}^2(x_i, x_j)) \cdot (\mu_{(b)lj} - \mu_{(b)li}), \quad (3.92)$$

где, в свою очередь, $\mu_{(b)li}$ — значение принадлежности i -го объекта l -му кластеру на b -м шаге вычислительного процесса.

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

$\hat{d}$ — порог расстояния между объектами;

Схема алгоритма:

- 1 Выбираются начальное разбиение $P = (A^1, \dots, A^c)$ на c нечетких классов, описываемое c непустыми функциями принадлежности, так что матрица начального разбиения $P = [\mu_{li}]$ имеет c строк и n столбцов; присваиваются значения $Q_{1(b-1)}^i(P) := 0$ и $b := 0$;
- 2 Вычисляется σ_b в соответствии с соотношением (3.86) и $\hat{d}_{(b)}$ в соответствии с соотношением (3.87); вычисляется значение $Q_{1(b)}^i(P)$ и сравнивается со значением $Q_{1(b-1)}^i(P)$: вычисляется некоторое пороговое значение $\varepsilon > 0$, если $Q_{1(b)}^i(P) \leq Q_{1(b-1)}^i(P)$, так что $|Q_{1(b)}^i(P) - Q_{1(b-1)}^i(P)| \leq \varepsilon$, то алгоритм прекращает работу;
- 3 Полагается $b_{(i)} := 1$ и $i := 1$; для i -го объекта вычисляются значения частных производных $\beta_{(b)i}(x_i)$, $\beta_{\max}(x_i)$ и $\beta_{\min}(x_i)$ по формулам (3.92), (3.90), (3.91) соответственно;
- 4 Если выполняется приближенное равенство $\beta_{\max}(x_i) \approx \beta_{\min}(x_i)$, то осуществляется переход на шаг 7; в противном случае осуществляется переход на шаг 5;
- 5 Для определения направления движения отыскивается наименьший положительный корень полинома $H(\lambda) = h_1\lambda^3 + h_2\lambda^2 + h_3\lambda + h_4$, где

$$h_1 = \sigma_b (1 - p(x_i)),$$

$$h_2 = \frac{3}{2} \sigma_b [(\mu_{fi} - \mu_{ki}) - (S_f - S_k)],$$

$$h_3 = \frac{1}{2} \left\{ \sigma_b \sum_{\substack{j=1 \\ j \neq i}}^n p(x_j) [(\mu_{fi} - \mu_{ki}) - (\mu_{fj} - \mu_{kj})]^2 + \sum_{\substack{j=1 \\ j \neq i}}^n p(x_j) [\sigma_b V_{(b)} - \hat{d}_{(b)}^2(x_i, x_j)] \right\}$$

$$h_4 = \frac{1}{4} [\beta_{\min}(x_i) - \beta_{\max}(x_i)],$$

где, в свою очередь,

$$\mu_{fi} = \max_{1 \leq l \leq c} \mu_{(b)li}, \quad (3.93)$$

$$\mu_{ki} = \min_{1 \leq l \leq c} \mu_{(b)li}, \quad (3.94)$$

причем относительные размеры S_f и S_k f -го и k -го нечетких кластеров, номера которых устанавливаются по формулам (3.93) и

(3.94), определяются по формуле $S_l = \sum_{j=1}^n p(x_j) \mu_{lj}$.

Поскольку $h_1 > 0$ и $h_2 < 0$, то в силу правила Декарта полином $H(\lambda)$ имеет один или три положительных корня; отыскивается наименьший положительный корень λ^+ полинома $H(\lambda)$ и сопоставляется с $\mu_{(b)li}$, так что $\lambda = \min(\mu_{(b)li}, \lambda^+)$;

- 6 Производится коррекция вектора степеней принадлежности i -го объекта в соответствии со следующим правилом:

$$\bar{\mu}_{fi} = \begin{cases} \mu_{fi} + \lambda, & \text{если } \mu_{fi} + \lambda \leq 1 \\ 1, & \text{иначе} \end{cases}, \quad (3.95)$$

$$\bar{\mu}_{ki} = \begin{cases} \mu_{ki} - \lambda, & \text{если } \mu_{ki} - \lambda \geq 0 \\ 0, & \text{иначе} \end{cases}, \quad (3.96)$$

где символами $\bar{\mu}_{fi}$ и $\bar{\mu}_{ki}$ обозначены измененные значения наибольшей и наименьшей степеней принадлежности i -го объекта; с учетом коррекции степеней принадлежности (3.95) и (3.96) производится также коррекция соответствующих S_f и S_k ; полагается $b_{(i)} := b_{(i)} + 1$;

- 7 Если просмотрены не все $x_i, i=1, \dots, n$, то полагается $i := i+1$ и осуществляется переход на шаг 3, в противном случае производится проверка изменения матрицы разбиения $P = [\mu_{li}]$: если имели место изменения $\mu_{li} \rightarrow \bar{\mu}_{li}, l=1, \dots, c, i=1, \dots, n$, то полагается $b := b+1$ и осуществляется переход на шаг 2, в противном случае полагается $P_{(b)} = P^*$ и алгоритм прекращает работу.

В работах [151], [152], [39] представлены результаты, развивающие идеи, изложенные в [149], [150].

3.2.2.2. Алгоритм Уиндхема (A-P algorithm) [186] минимизирует функционал $Q_2^i(P)$; поскольку результатом работы алгоритма является не только матрица разбиения $P_{\text{сxn}} = [\mu_{li}]$, но и матрица весов прототипов $K_{\text{сxn}} = [v_{li}]$, то $Q_2^i(P)$ будет записываться в виде $Q_2^i(P, K)$. Таким образом, алгоритм Уиндхема находит решение оптимизационной задачи $Q^i(P) \rightarrow \min$ в следующем виде:

$$P^* = \arg \min_P \left\{ \begin{array}{l} Q_2^i(P, K) : P = (A^1, \dots, A^c), A^l = [(\mu_{l1}, \dots, \mu_{ln}), (v_{l1}, \dots, v_{ln})], 0 \leq \mu_{li} \leq 1, \\ 0 \leq v_{lj} \leq 1, \sum_{l=1}^c \mu_{li} = 1, \sum_{j=1}^n v_{lj} = 1, \sum_{i=1}^n \mu_{li} > 0, \sum_{l=1}^c v_{lj} > 0, i, j = 1, \dots, n, l = 1, \dots, c \end{array} \right\}.$$

В силу особенностей функционала $Q_2^i(P, K)$ алгоритм нахождения

оптимального разбиения P^* несколько отличается от остальных нечетких оптимизационных кластер-процедур.

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

Схема алгоритма:

- 1 Выбирается начальное разбиение $P_{(0)} = (A_{(0)}^1, \dots, A_{(0)}^c)$ на c нечетких классов, так что матрица начального разбиения $P = [\mu_{ij}]$ имеет c строк и n столбцов; $b := 1$;
- 2 Полагается $P_{(b)} := P_{(0)}$ и вычисляется матрица весов прототипов $K_{(0)}$ в соответствии с соотношением (3.32);
- 3 Полагается $K_{(b+1)} := K_{(b)}$ и вычисляется матрица разбиения $P_{(b)}$ в соответствии с соотношением (3.31);
- 4 Вычисляется матрица весов прототипов $K_{(b)}$ на основании матрицы $P_{(b)}$ в соответствии с соотношением (3.32);
- 5 Выбирается некоторая мера отклонения $\varepsilon > 0$ и сравниваются значения $Q_2^i(P_{(b-1)}, K_{(b-1)})$ и $Q_2^i(P_{(b)}, K_{(b)})$: если $Q_2^i(P_{(b-1)}, K_{(b-1)}) - Q_2^i(P_{(b)}, K_{(b)}) < \varepsilon$, то $P^* := P_{(b)}$, $K^* := K_{(b)}$ и алгоритм прекращает работу, в противном случае полагается $b := b + 1$ и осуществляется переход на шаг 3.

При фиксированной матрице весов прототипов K матрица разбиения P , строящаяся в соответствии с формулой (3.31), минимизирует $Q_2^i(P_{(b)}, K)$ по всем матрицам разбиения $P_{(b)}$, так что выполняется $Q_2^i(P_{(b)}, K_{(b-1)}) \leq Q_2^i(P_{(b-1)}, K_{(b-1)})$. Аналогично $Q_2^i(P_{(b)}, K)$ минимизируется матрицей K , строящейся по формуле (3.32), так что $Q_2^i(P_{(b)}, K_{(b)}) \leq Q_2^i(P_{(b)}, K_{(b-1)})$; таким образом, алгоритм уменьшает значение целевой функции $Q_2^i(P, K)$ на каждой итерации. Учитывая это обстоятельство, а также то, что функция $Q_2^i(P, K)$ неотрицательна, следует, что последовательность $\{Q_2^i(P_{(b)}, K_{(b)})\}$ сходится, так что при любом $\varepsilon > 0$ алгоритм остановится после конечного числа шагов.

Относительное уменьшение целевой функции $Q_2^i(P, K)$ также может быть использовано для прекращения работы алгоритма, для чего на шаге 5 алгоритма условие $Q_2^i(P_{(b-1)}, K_{(b-1)}) - Q_2^i(P_{(b)}, K_{(b)}) < \varepsilon$ может быть заменено выражением вида $(Q_2^i(P_{(b-1)}, K_{(b-1)}) - Q_2^i(P_{(b)}, K_{(b)})) / Q_2^i(P_{(b)}, K_{(b)}) < \varepsilon$. Такая нормализация

обладает тем преимуществом, что выбор порога ε не зависит от величины целевой функции. К примеру, выбор порога $\varepsilon = 0.0001$ интерпретируется как останов вычислений в случае, когда целевая функция $Q_2^l(P, K)$ уменьшается менее чем на 0.01 процента.

Вопросы, связанные с выбором начального разбиения на шаге 1 А-Р алгоритма, также подробно рассматриваются в работе [186, с.165-166].

3.2.2.3. Алгоритм Рубенса (MND2 algorithm) [148] минимизирует критерий $Q_3^l(P)$, отыскивая решение

$$P^* = \arg \min_P \left\{ Q_3^l(P) : P = (A^1, \dots, A^c), A^l = (\mu_{li}, \dots, \mu_{ln}), 0 \leq \mu_{li} \leq 1, \right. \\ \left. \sum_{i=1}^c \mu_{li} = 1, \sum_{i=1}^n \mu_{li} > 0, i = 1, \dots, n, l = 1, \dots, c \right\}.$$

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

Схема алгоритма:

1 Выбирается начальное разбиение $P = (A^1, \dots, A^c)$ на c нечетких классов, описываемое c непустыми функциями принадлежности, так что матрица начального разбиения $P = [\mu_{li}]$ имеет c строк и n столбцов;

2 Вычисляется функция расстояния $D(x_i, l) = \sum_{j=1}^n \mu_{lj}^2 d(x_i, x_j)$;

3 Решается следующая задача квадратичного программирования:

$$\sum_{i=1}^c \sum_{l=1}^c \mu_{li}^2 D(x_i, l) \rightarrow \min_{\mu_{li}} \text{ при } \mu_{li} \geq 0, \sum_{i=1}^c \mu_{li} = 1, \text{ с использованием множителей Лагранжа:}$$

$$\sum_{i=1}^c \sum_{l=1}^n \mu_{li}^2 D(x_i, l) + \sum_{i=1}^n \lambda(x_i) \left[\sum_{l=1}^c \mu_{li} - 1 \right] \rightarrow \min_{\mu_{li}},$$

что приводит к

$$\frac{\partial}{\partial \mu_{li}} = 0 : \sum_{i=1}^c \mu_{li} = 1;$$

$$\frac{\partial}{\partial \lambda(x_i)} = 0 : 2\mu_{li} D(x_i, l) + \lambda(x_i) = 0;$$

$$\mu_{li} = \left\{ D(x_i, l) \sum_{a=1}^c D^{-1}(x_i, a) \right\}, \mu_{li} \geq 0;$$

- 4 В зависимости от результата сравнения полученного разбиения с предыдущим разбиением осуществляется переход на шаг 2 или останов алгоритма.

3.2.2.4. Алгоритм Беждека — Данна (Fuzzy ISODATA algorithm, FCM algorithm) [83], [85], [53], [59], [55], [61] в приведенной ниже версии минимизирует критерий $Q_1^H(P)$ в виде (3.46), так что решение задачи классификации находится в следующем виде:

$$P^* = \arg \min_P \left\{ Q_1^H(P) : P = (A^1, \dots, A^c), A^l = (\mu_{l1}, \dots, \mu_{ln}), 0 \leq \mu_{li} \leq 1, \right. \\ \left. \sum_{l=1}^c \mu_{li} = 1, \sum_{i=1}^n \mu_{li} > 0, i = 1, \dots, n, l = 1, \dots, c \right\}.$$

Параметры алгоритма:

- c — число нечетких кластеров в искомом разбиении P^* ;
 γ — показатель нечеткости классификации, $1 < \gamma < \infty$;

Схема алгоритма:

- 1 Выбирается начальное разбиение $P_{(0)} = (A_{(0)}^1, \dots, A_{(0)}^c)$ на c нечетких классов, описываемое c непустыми функциями принадлежности, которое представляет собой массив

$$\{\mu_{(0)1}, \dots, \mu_{(0)n}\}, \mu_{(0)i} = (\mu_{(0)i1}, \dots, \mu_{(0)ci}), \mu_{(0)li} \geq 0, \sum_{l=1}^c \mu_{(0)li} = 1 \text{ из } n \text{ } c\text{-}$$

мерных столбцов, для всех $i = 1, \dots, n$, так что полученная матрица начального разбиения $P_{(0)} = [\mu_{(0)li}]$ имеет c строк и n столбцов;

$b := 0$;

- 2 Пусть построено некоторое b -е разбиение $P_{(b)}$ в виде массива

$$\{\mu_{(b)1}, \dots, \mu_{(b)n}\}, \mu_{(b)i} = (\mu_{(b)i1}, \dots, \mu_{(b)ci}), \mu_{(b)li} \geq 0, \sum_{l=1}^c \mu_{(b)li} = 1 \text{ из } n \text{ } c\text{-}$$

мерных столбцов; вычисляется набор центров $\tau_{(b)}^1, \dots, \tau_{(b)}^c$ в соответ-

ствии с формулой $\tau_{(b)}^l = \sum_{i=1}^n \mu_{(b)li}^\gamma \cdot x_i / \sum_{i=1}^n \mu_{(b)li}^\gamma, l = 1, \dots, c$;

- 3 Строится $b+1$ -е разбиение $P_{(b+1)}$ в виде массива

$$\mu_{(b+1)1}, \dots, \mu_{(b+1)n}, \mu_{(b+1)i} = (\mu_{(b+1)i1}, \dots, \mu_{(b+1)ci}), \mu_{(b+1)li} \geq 0, \sum_{l=1}^c \mu_{(b+1)li} = 1 \text{ из } n$$

c -мерных столбцов, порождаемое набором центров $\tau_{(b)}^1, \dots, \tau_{(b)}^c$, где

$$\mu_{(b+1)i} = \arg \min \left\{ \sum_{l=1}^c \mu_l^\gamma \|x_i - \tau_{(b)}^l\|^2 : \mu = \mu_1, \dots, \mu_c, \mu_l \geq 0, \sum_{l=1}^c \mu_l = 1 \right\}, \text{ то}$$

есть если $I_{(b)}(x) = \{l | 1 \leq l \leq c : \|x_i - \tau_{(b)}^l\| = 0\}$, то

$$I_{(b)}(x) = \emptyset, \mu_{(b+1)li} = \left(\sum_{a=1}^c \left(\frac{\|x_i - \tau_{(b)}^l\|}{\|x_i - \tau_{(b)}^a\|} \right)^{\frac{2}{\gamma-1}} \right)^{-1},$$

$$I_{(b)}(x) \neq \emptyset, \mu_{(b+1)li} = 0, l \neq I_{(b)}(x); \sum_{l \in I_{(b)}(x)} \mu_{(b+1)li} = 1;$$

- 4 Вычисляется некоторое пороговое значение $\varepsilon > 0$ и производится сравнение $P_{(b)}$ и $P_{(b+1)}$ по правилу: $\|P_{(b)} - P_{(b+1)}\| = \max_{l,i} |\mu_{(b)li} - \mu_{(b+1)li}|$; если $\|P_{(b)} - P_{(b+1)}\| < \varepsilon$, то $P_{(b)} = P^*$ и алгоритм заканчивает работу, в противном случае полагается $b := b + 1$ и осуществляется переход на шаг 2.

3.2.2.5. Алгоритм Педрича [140], [144] в предлагаемом изложении минимизирует функционал $Q_2^n(P)$ в виде (3.56), так что результатом работы процедуры будет следующее решение задачи классификации:

$$P^* = \arg \min_P \left\{ Q_2^n(P) : P = (A^1, \dots, A^c), A^l = (\mu_{l1}, \dots, \mu_{ln}), 0 \leq \mu_{li} \leq 1, \right. \\ \left. \sum_{l=1}^c \mu_{li} = 1, \sum_{i=1}^n \mu_{li} > 0, i = 1, \dots, n, l = 1, \dots, c \right\}.$$

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

Схема алгоритма:

- 1 Выбирается начальное разбиение $P_{(0)} = (A_{(0)}^1, \dots, A_{(0)}^c)$ на c нечетких классов, описываемое c непустыми функциями принадлежности, так что полученная матрица начального разбиения $P_{(0)} = [\mu_{(0)li}]$ имеет c строк и n столбцов; $b := 1$;
- 2 Вычисляется набор центров $\tau_{(b)}^1, \dots, \tau_{(b)}^c$ в соответствии с формулой
$$\tau_{(b)}^l = \sum_{i=1}^n \mu_{(b-1)li}^2 \cdot x_i / \sum_{i=1}^n \mu_{(b-1)li}^2, l = 1, \dots, c;$$
- 3 Строится разбиение $P_{(b)}$ в соответствии с формулой

$$\mu_{(b)li} = \frac{2 - \sum_{a=1}^c y_{ai} s_i}{2 \sum_{a=1}^c \left(\frac{d(x_i, \tau_{(b)}^l)}{d(x_i, \tau_{(b)}^a)} \right)^2} + \frac{1}{2} y_{li} s_i;$$

- 4 Вычисляется некоторое пороговое значение $\varepsilon > 0$ и производится сравнение $P_{(b)}$ и $P_{(b-1)}$ по правилу $\|P_{(b)} - P_{(b-1)}\| = \max_{l,i} |\mu_{(b)li} - \mu_{(b-1)li}|$. Если $\|P_{(b)} - P_{(b-1)}\| < \varepsilon$, то $P_{(b)} = P^*$ и алгоритм заканчивает работу; в противном случае полагается $b := b + 1$ и осуществляется переход на шаг 2.

Вычислительные эксперименты показали очень хорошие результаты метода в сравнении с различными модификациями алгоритма Беждека — Данна, особенно при использовании расстояния Махаланобиса; в частности, успешно были расклассифицированы данные, известные как «крест Густафсона» [92], [140, с.16]. Дальнейшее развитие подхода В. Педрича рассматривается в его работах [142], [145].

3.2.2.6. Алгоритм Даве — Сена (FRC algorithm) [73] использует вспомогательные значения v_{li} для построения значений принадлежности $\mu_{li}, l = 1, \dots, c, i = 1, \dots, n$ и минимизирует критерий $Q'_5(P)$, отыскивая решение

$$P^* = \arg \min_P \left\{ \begin{array}{l} Q'_5(P) : P = (A^1, \dots, A^c), A^l = (\mu_{l1}, \dots, \mu_{ln}), 0 \leq \mu_{li} \leq 1, \\ \sum_{i=1}^c \mu_{li} = 1, \sum_{i=1}^n \mu_{li} > 0, i = 1, \dots, n, l = 1, \dots, c \end{array} \right\}.$$

Параметры алгоритма:

- c — число нечетких кластеров в искомом разбиении P^* ;
 γ — показатель нечеткости классификации, $1 < \gamma < \infty$;

Схема алгоритма:

- 1 Выбирается начальное разбиение $P = (A^1, \dots, A^c)$ на c нечетких классов, так что матрица начального разбиения $P = [\mu_{li}]$ имеет c строк и n столбцов; полагается $b := 0$;
- 2 Полагается $i := 1$;
- 3 Вычисляются значения v_{li} в соответствии с соотношением

$$v_{li} = \frac{\gamma \cdot \sum_{j=1}^n \mu_{lj}^\gamma d(x_i, x_j)}{\sum_{j=1}^n \mu_{lj}^\gamma} - \frac{\gamma \cdot \sum_{k=1}^n \sum_{j=1}^n \mu_{lj}^\gamma \mu_{ik}^\gamma d(x_j, x_k)}{2 \left(\sum_{j=1}^n \mu_{lj}^\gamma \right)^2}$$

для всех $1 \leq l \leq c$ с использованием вновь вычисленных значений принадлежности μ_{lj} , если $j < i$, и предыдущих значений принадлежности μ_{lj} при $j \geq i$;

- 4 Производится пересчет значений принадлежности в матрице $P = [\mu_{ij}]$ в соответствии с соотношением

$$\mu_{ij} = \frac{\left(\frac{1}{v_{ij}}\right)^{1/(r-1)}}{\sum_{a=1}^c \left(\frac{1}{v_{ai}}\right)^{1/(r-1)}};$$

- 5 Если $i < n$ то полагается $i := i + 1$ и осуществляется переход на шаг 3;
- 6 Полагается $b := b + 1$; если разница между полученным и предыдущим разбиениями не превышает некоторой выбранной меры отклонения $\varepsilon > 0$ или достигнуто заданное число итераций $b_{\max}$, осуществляется останов алгоритма; в противном случае осуществляется переход на шаг 3.

3.2.2.7. Другие алгоритмы оптимизационного подхода подробно рассматриваются в работах [46], [64], [85], [101], [108], [167], [185] и, в основном, представляют собой различные версии рассмотренных выше нечетких оптимизационных кластер-процедур. Дж. Беждеком [59] рассматриваются различные модификации критерия $Q_1''(P)$ и соответствующие им алгоритмы нахождения оптимума, такие, к примеру, как FCV алгоритм (fuzzy c-varieties). Алгоритм, предложенный Д. Е. Густафсоном и В. С. Кесселем [92] и минимизирующий критерий $Q_1''(P)$, где в качестве функции $d(x_i, \tau^i)$ используется квадрат расстояния Махаланобиса (3.47), известен в литературе как GK алгоритм [109].

Некоторые подходы используют методы нечеткой математики для усовершенствования статистических процедур; наиболее интересным с этой точки зрения представляется метод, предложенный Я. В. Овсинским и С. Задрожным в работе [133], совершенствующий процедуру, отыскивающую субоптимум целевой функции [132], на основе шести предложенных правил.

Ниже рассмотрены некоторые малоизвестные процедуры, представляющие интерес с теоретической точки зрения, а также ряд модификаций рассмотренных выше методов.

Алгоритм Распини, минимизирующий один из функционалов $Q_{1(1)}'(P)$ или $Q_{1(2)}'(P)$, предложен в работе [150, с.324] и используется в описанном выше алгоритме минимизации критерия $Q_1'(P)$, так что его

рассмотрение оказывается нелишним с теоретической точки зрения. Однако перед изложением общей схемы процедуры следует привести формулировки двух теорем, используемых в ней.

Как и функционал $Q_1^l(P)$, функционалы $Q_{1(1)}^l(P)$ и $Q_{1(2)}^l(P)$ представляют собой дифференцируемые функции, частные производные которых находятся по формулам (3.70), (3.71), (3.72):

$$\beta_i(x_i) = -\frac{\partial Q_{1(z)}^l(P)}{\partial \mu_{li}}, l=1, \dots, c, i=1, \dots, n, z=1, 2, \quad \beta_{\max}(x_i) = \max_{(i)} \beta_i(x_i) \quad \text{и} \\ \beta_{\min}(x_i) = \min_{(i)} \beta_i(x_i).$$

Теорема 3.1. Если для некоторого разбиения $P^* = (A^1, \dots, A^c)$, описываемого функциями принадлежности $\mu_{li}, l=1, \dots, c, i=1, \dots, n$, выполняется равенство $\beta_{\max}(x_i) = \beta_{\min}(x_i), 1 \leq i \leq n$, то данное разбиение P^* определяет стационарную точку функции $Q_{1(1)}^l(P)$ или $Q_{1(2)}^l(P)$, то есть для всех возможных направлений $k_l(x_i)$, таких, что

$$\sum_{l=1}^c k_l(x_i) = 0, \sum_{l=1}^c k_l^2(x_i) = 1, 1 \leq i \leq n, \text{ и } k_l(x_i) \geq 0, l \notin J_i = \{l, 1 \leq l \leq c : \mu_{li} > 0\},$$

$$\text{выполняется } \sum_{l=1}^c \sum_{i=1}^n k_l(x_i) \beta_l(x_i) \leq 0.$$

Теорема 3.2. Если при условиях предыдущей теоремы для некоторого $j, 1 \leq j \leq n$ выполняется $\beta_{\max}(x_j) > \beta_{\min}(x_j)$, то выбираются $k_{\max}(x_j) = \sqrt{2}/2$ и $k_{\min}(x_j) = -\sqrt{2}/2$, а также $k_l(x_i) = 0, 1 \leq l \leq c, 1 \leq i \leq n$ такие, что $i \neq j$ и индекс l не совпадает с индексами $\max$ и $\min$ нечеткого кластера, и $\sum_{l=1}^c \sum_{i=1}^n k_l(x_i) \beta_l(x_i) > 0$.

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

Схема алгоритма:

1. Для некоторого разбиения $P = (A^1, \dots, A^c)$, описываемого функциями принадлежности $\mu_{li}, 1 \leq l \leq c, 1 \leq i \leq n$, некоторая точка x_i фиксируется, так что $i = j$, и для j -го объекта вычисляются значения частных производных $\beta_l(x_j), 1 \leq l \leq c$;
2. Из множества значений частных производных $\{\beta_l(x_j), 1 \leq l \leq c\}$ вычисляются $\beta_{\max}(x_j)$ и $\beta_{\min}(x_j)$;

3. Если выполняется равенство $\beta_{\max}(x_j) = \beta_{\min}(x_j)$, то шаги 1 и 2 повторяются для всех $j+1, \dots, n$;
4. Если для некоторой j -й точки, $1 \leq j \leq n$, выполняется $\beta_{\max}(x_j) \neq \beta_{\min}(x_j)$, то выполняются вычисления в соответствии с условием теоремы 3.2;
5. Если для всех $j, i=1 \leq j \leq n$ выполняется равенство $\beta_{\max}(x_j) = \beta_{\min}(x_j)$, то найденная точка является оптимальной.

Теоремы 3.1 и 3.2 показывают, что предложенный градиентный метод сходится при указанных ограничениях. Важность рассмотренного метода заключается в том, что минимизация функционала достигается изменением значений принадлежности μ_j в некоторой точке $x_j \in X$ при фиксированных значениях $\mu_i, 1 \leq i \leq n, i \neq j$, а значение j последовательно изменяется на каждом шаге. Усовершенствованный вариант процедуры предусматривает цикл, повторяющий шаг 4 процедуры для фиксированной точки $x_j \in X$ до тех пор, пока не выполняется равенство $\beta_{\max}(x_j) = \beta_{\min}(x_j)$, после чего происходит изменение значения j .

В работе [150, с.325-329] также рассматривается процедура, минимизирующая функционалы $Q'_{1(1)}(P)$ и $Q'_{1(2)}(P)$ одновременно. Данная процедура устанавливает связь между рассмотренной процедурой минимизации одного из функционалов $Q'_{1(1)}(P)$ или $Q'_{1(2)}(P)$, с одной стороны, и процедурой минимизации функционала $Q'_1(P)$ — с другой, и, хотя и представляет интерес с теоретической точки зрения, ее детальное рассмотрение из-за достаточно высокой сложности и громоздкости нецелесообразно.

Алгоритм расстановки меток [72, с.173-174] применяется для присвоения названий классов центрам нечетких кластеров, что в значительной степени облегчает интерпретацию результатов. Обычно хотя бы для части объектов исследуемой совокупности точно известно, какой объект принадлежит какому классу. Данная информация может послужить основой процесса присвоения названий классов центрам нечетких кластеров, именуемым расстановкой меток (labeling). В случае, когда априорная информация о принадлежности некоторых объектов классам отсутствует, то она может быть получена уже после процедуры классификации объектов, в данном случае методом нечетких s -средних.

Пусть в результате работы алгоритма Беждека — Данна построено нечеткое разбиение на c классов и результат представлен в виде соответствующей матрицы $P_{c \times n} = [\mu_{li}]$, а $X' = \{x'_i\}, i = 1, \dots, c$ — некоторое подмножество классифицированных объектов, $X' \subset X, \text{card}(X') = c$, так что $x'_1 \in A^1, x'_2 \in A^2, \dots, x'_c \in A^c$; инициализируется матрица меток $L_{c \times c} = [L_{il}]$.

Параметры алгоритма:

c — число нечетких кластеров в разбиении P^* ;

γ — показатель нечеткости классификации, $1 < \gamma < \infty$;

Схема алгоритма:

- 1 Всем элементам матрицы меток $L_{c \times c} = [L_{il}]$ присваивается значение 0; $i := 1$;
- 2 Вычисляется значение функции принадлежности μ'_{li} объекта x'_i для всех центров кластеров $\tau^l, l = 1, \dots, c$ в соответствии со следующей формулой:

$$\mu'_{li} = \frac{1}{\sum_{a=1}^c \left(\frac{\|x'_i - \tau^l\|}{\|x'_i - \tau^a\|} \right)^{\frac{2}{\gamma-1}}};$$

- 3 Полагается $L_{il} := L_{il} + \mu'_{li}$ для всех $l = 1, \dots, c$;
- 4 Если $i < c$, то полагается $i := i + 1$ и осуществляется переход на шаг 2;
- 5 Определяется метка $L(l) = \max_{i=1, \dots, c} \{L_{il}\}$ для всех $l = 1, \dots, c$;
- 6 Производится присвоение меток $L(l), l = 1, \dots, c$ центрам кластеров $\tau^l, l = 1, \dots, c$.

Следует указать, что процедура расстановки меток терпит неудачу, если метка присвоена двум или более классам, что означает, что центр кластера не найден для всех меток классов. Иногда может случиться, что максимальное значение матрицы меток на шаге 5 однозначно не устанавливается. В обеих указанных ситуациях алгоритм останавливает работу и выдает сообщение об ошибке.

Модификация Либерта — Рубенса MND2 алгоритма [119, с.418], в которой полагается, что $\mu_{li} = \mu_{lj}$ при $x_i = x_j$, незначительно отличается от приведенной выше версии метода MND2.

Параметры алгоритма:

c — число нечетких кластеров в искомом разбиении P^* ;

Схема алгоритма:

1 Выбирается начальное разбиение $P = (A^1, \dots, A^c)$ на c нечетких классов, описываемое c непустыми функциями принадлежности, так что матрица начального разбиения $P = [\mu_{ij}]$ имеет c строк и n столбцов;

2 Полагается $i := 0$;

3 Полагается $i := i + 1$; вычисляется функция расстояния

$$D(x_i, l) = \sum_{j=1}^n \mu_{ij}^2 d(x_i, x_j) \text{ для } l = 1, \dots, c \text{ и решается следующая задача}$$

квадратичного программирования: $\sum_{l=1}^c \sum_{i=1}^c \mu_{li}^2 D(x_i, l) \rightarrow \min_{\mu_l}$ при

$\mu_{li} \geq 0, \sum_{l=1}^c \mu_{li} = 1$, с использованием множителей Лагранжа:

$$\sum_{l=1}^c \sum_{i=1}^n \mu_{li}^2 D(x_i, l) + \lambda \left[\sum_{l=1}^c \mu_{li} - 1 \right] \rightarrow \min_{\mu_l},$$

что приводит к

$$\frac{\partial}{\partial \mu_{li}} = 0: \sum_{l=1}^c \mu_{li} = 1;$$

$$\frac{\partial}{\partial \lambda} = 0: 4\mu_{li} D(x_i, l) + \lambda = 0;$$

$$\mu_{li}^{-1} = D(x_i, l) \sum_{a=1}^c D^{-1}(x_i, a), \mu_{li} \geq 0;$$

4 Полагается $\mu_{ij} = \mu_{li}$ для $l = 1, \dots, c$ и $x_j = x_i$;

5 Если $i < n$, то осуществляется переход на шаг 3;

6 В зависимости от результата сравнения полученного разбиения с предыдущим разбиением осуществляется переход на шаг 2 или останов алгоритма.

М. П. Уиндхем отмечал, что «алгоритм Рубенса основан на очень простом оптимизационном критерии, но первоначально сформулированная численная процедура неустойчива. Модификации, описанные Либбертом и Рубенсом, приводят к более стабильной технике, но если типы явно неотличимы друг от друга, полезные результаты могут не быть получены» [186, с.159].

Модификация Даве — Сена FCM алгоритма (R-FCM algorithm) [73, с.716] минимизирует функционал $Q_{1R}^n(P)$ (3.69) и практически не отличается от схемы FCM алгоритма, представляя собой его робастную версию: формула для вычисления центров классов остается

ся неизменной, тогда как значения принадлежностей вычисляются по формуле (3.70).

Модификация Даве — Сена MND2 алгоритма (R-MND2 algorithm) [73, с.716] минимизирует функционал $Q'_{3R}(P)$ (3.74) и отличается от рассмотренной выше схемы MND2 алгоритма тем, что введение класса «выбросов» A^* обуславливает определение для него функции расстояния в виде $D(x_i, *) = \sum_{j=1}^n \mu_{*j}^2 \delta = \delta N_*$, где $N_* = \sum_{j=1}^n \mu_{*j}^2$, а значения принадлежностей вычисляются по формуле

$$\mu_{ii} = \frac{\left(\frac{1}{D(x_i, l)} \right)}{\sum_{a=1}^c \left(\frac{1}{D(x_i, a)} \right) + \left(\frac{1}{\delta N_*} \right)}.$$

Модификация Даве — Сена FRC алгоритма (R-FRC algorithm) [73, с.716] является робастной версией FRC алгоритма и минимизирует функционал $Q'_{5R}(P)$ (3.75), так что схема R-FRC алгоритма отличается от схемы FRC алгоритма тем, что значения принадлежностей вычисляются в соответствии с формулой

$$\mu_{ii} = \frac{\left(\frac{1}{v_{ii}} \right)^{1/(r-1)}}{\sum_{a=1}^c \left(\frac{1}{v_{ai}} \right)^{1/(r-1)} + \left(\frac{2}{\gamma \delta} \right)^{1/(r-1)}}, \text{ где, в свою очередь, значения } v_{ii} \text{ вычисляются, как и на шаге 3 FRC алгоритма.}$$

В работе [73] представлена также робастная версия функционала, предложенного Р. Дж. Гэтеем, Дж. В. Девенпортом и Дж. Беждеком, так что минимизирующий робастную версию функционала Р. Дж. Гэтеева, Дж. В. Девенпорта и Дж. Беждека алгоритм получил название R-RFCM алгоритма. Кроме того, Р. Н. Даве и С. Сеном рассматривается версия FRC алгоритма, отыскивающая оптимальное разбиение при условии, что для расстояния $d(x_i, x_j)$ между объектами исследуемой совокупности $X = \{x_1, \dots, x_n\}$ имеет место $d(x_i, x_j) \neq \|x_i - x_j\|^2, i, j = 1, \dots, n$, получившая название NE-FRC алгоритма, а также робастная версия NE-FRC алгоритма (R-NE-FRC algorithm).

Модификация Кеймака — Сетнеса GK алгоритма (E-GK algorithm) [109, с.708-709] минимизирует критерий $Q_{1E}''(P)$ (3.76), где в качестве функции расстояния используется квадрат расстояния Махаланобиса (3.47).

Параметры алгоритма:

c — число нечетких кластеров в разбиении;

γ — показатель нечеткости классификации, $1 < \gamma < \infty$;

Схема алгоритма:

- 1 Выбирается начальное число кластеров $1 < c^{(0)} < n$, меры отклонения $\varepsilon_1, \varepsilon_2 > 0$ и строится начальное разбиение $P_{(0)} = (A_{(0)}^1, \dots, A_{(0)}^{c^{(0)}})$ на $c^{(0)}$ нечетких классов, описываемое $c^{(0)}$ непустыми функциями принадлежности, так что полученная матрица начального разбиения $P_{(0)} = [\mu_{(0)li}]$ имеет $c^{(0)}$ строк и n столбцов; полагается $\chi^{(0)} := 1$ и $S_{i_0 f_0}^{(0)} = 1$;
- 2 Полагается $b := 1$;
- 3 Пусть построено некоторое b -е разбиение $P_{(b)}$ в виде массива $\{\mu_{(b)1}, \dots, \mu_{(b)n}\}$, $\mu_{(b)i} = (\mu_{(b)1i}, \dots, \mu_{(b)ci})$, $\mu_{(b)li} \geq 0$, $\sum_{l=1}^c \mu_{(b)li} = 1$ из n столбцов; вычисляется набор центров $\tau_{(b)}^1, \dots, \tau_{(b)}^c$ в соответствии с формулой $\tau_{(b)}^l = \sum_{i=1}^n \mu_{(b-1)li}^\gamma \cdot x_i / \sum_{i=1}^n \mu_{(b-1)li}^\gamma$, $l = 1, \dots, c^{(b-1)}$;
- 4 На основании ковариационной матрицы $U_l = \left(\sum_{i=1}^n \mu_{(b-1)li}^\gamma (x_i - \tau_{(b)}^l)(x_i - \tau_{(b)}^l)^T \right) / \left(\sum_{i=1}^n \mu_{(b-1)li}^\gamma \right)$ вычисляется радиус прототипов кластеров $R_l = \chi^{(b-1)} \frac{\sqrt{|U_l|}^{1/m}}{c^{(b-1)}}$, $l = 1, \dots, c^{(b-1)}$;
- 5 Вычисляются расстояния до объемных прототипов кластеров: $\tilde{d}(x_i, \tilde{\tau}^l) = \max \left(0, \frac{(x_i - \tau_{(b)}^l)^T U_l^{-1} (x_i - \tau_{(b)}^l)}{|U_l|^{1/m}} - R_l^2 \right)$, $l = 1, \dots, c^{(b-1)}$, $i = 1, \dots, n$;
- 6 Производится пересчет значений принадлежности в матрице разбиения $P_{(b)} = [\mu_{(b)li}]$ в соответствии со следующим правилом: для $i = 1, \dots, n$ полагается $I(x_i) = \{l \mid \tilde{d}(x_i, \tilde{\tau}^l) = 0\}$, так что

$$I(x_i) = \emptyset, \mu_{(b)li} = \frac{1}{\sum_{a=1}^{c^{(b-1)}} \left(\frac{\tilde{d}(x_i, \tilde{\tau}^l)}{\tilde{d}(x_i, \tilde{\tau}^a)} \right)^{1/(\gamma-1)}}, 1 \leq l \leq c^{(b-1)},$$

$$I(x_i) \neq \emptyset, \mu_{(b)li} = \begin{cases} 0, & \tilde{d}(x_i, \tilde{\tau}^l) > 0 \\ \frac{1}{|I(x_i)|}, & \tilde{d}(x_i, \tilde{\tau}^l) = 0 \end{cases}, 1 \leq l \leq c^{(b-1)};$$

7 Выбирается пара наиболее близких кластеров:

$$S_{lf}^{(b)} = \frac{\sum_{i=1}^n \min(\mu_{(b)li}, \mu_{(b)fi})}{\min\left(\sum_{i=1}^n \mu_{(b)li}, \sum_{i=1}^n \mu_{(b)fi}\right)}, 1 \leq l, f \leq c^{(b-1)}, (A_*^l, A_*^f) = \arg \max_{l, f \in I} (S_{lf}^{(b)});$$

8 Производится объединение наиболее близких кластеров в соответствии с правилом: при $|S_{l_*f_*}^{(b)} - S_{l_*f_*}^{(b-1)}| < \varepsilon_1$ полагается $\alpha^{(b)} = 1/(c^{(b-1)} - 1)$; если $S_{l_*f_*}^{(b)} > \alpha^{(b)}$, то $\mu_{l_*i}^{(b)} := \mu_{l_*i}^{(b-1)} + \mu_{f_*i}^{(b-1)}, i = 1, \dots, n$, из матрицы $P_{(b)} = [\mu_{(b)li}]$ удаляется строка, соответствующая кластеру A_*^f , и полагается $c^{(b)} = c^{(b-1)} - 1$, иначе $\chi^{(b)} = \min(c^{(b-1)}, \chi^{(b-1)} + 1)$;

9 Производится сравнение $P_{(b)}$ и $P_{(b-1)}$ по правилу: если $\|P_{(b)} - P_{(b-1)}\| < \varepsilon_2$, то $P_{(b)} = P^*$ и алгоритм заканчивает работу; в противном случае полагается $b := b + 1$ и осуществляется переход на шаг 3.

Модификация Кеймака — Сетнеса FCM алгоритма (E-FCM algorithm) [109, с.708-709] минимизирует критерий $Q_{1E}^H(P)$ (3.76), где в качестве функции расстояния используется квадрат евклидова расстояния (3.45), так что E-FCM алгоритм отличается от E-GK алгоритма только способом вычисления расстояния на шаге 5: $\tilde{d}(x_i, \tilde{\tau}^l) = \max(0, (x_i - \tau_{(b)}^l)^T (x_i - \tau_{(b)}^l) - R_l^2), l = 1, \dots, c^{(b-1)}, i = 1, \dots, n$.

Результаты вычислительных экспериментов, представленные У. Кеймаком и М. Сетнесом [109, с.709], демонстрируют высокую эффективность расширенных версий нечетких кластер-процедур оптимизационного направления по сравнению с обычными нечеткими оптимизационными методами кластер-анализа. При проведении экспериментов У. Кеймаком и М. Сетнесом полагалось $\gamma = 2$, $\varepsilon_1 = 0.01$ и $\varepsilon_2 = 0.001$.

В заключение рассмотрения оптимизационных алгоритмов нечеткого подхода к решению задачи кластер-анализа следует отметить общую особенность всех рассмотренных кластер-процедур [109, с.705]: результаты работы оптимизационных алгоритмов зависят от начального разбиения, которое инициализируется, как правило, случайным образом, так что при неудовлетворительном начальном разбиении могут быть получены некорректные результаты. В связи с этим У. Кеймак и М. Сетнес отмечают, что для получения корректных результатов целесообразно проведение ряда экспериментов с различными начальными разбиениями [109, с.705].

3.3. Иерархические методы

3.3.1. Основные понятия иерархического направления решения нечеткой модификации задачи автоматической классификации

Методы иерархического направления нечеткого подхода к решению задачи автоматической классификации образуют самую немногочисленную группу и представлены двумя процедурами. Иерархические методы ставят своей задачей построение иерархии на классифицируемой совокупности объектов $X = \{x_1, \dots, x_n\}$, характеризующей нечеткостью, и наглядное представление ее стратификационной структуры. Построение нечеткой иерархии на множестве объектов $X = \{x_1, \dots, x_n\}$ означает, таким образом, выявление принципа объединения объектов в нечеткие кластеры в случае агломеративных алгоритмов либо, напротив, принципа расслоения классифицируемой совокупности объектов на нечеткие кластеры, вплоть до отдельных объектов, если применяются дивизимные кластер-процедуры.

Алгоритм, предложенный Д. Ватадой, Х. Танакой и К. Асаи [182], строит нечеткое отношение $\tilde{T}$ на множестве объектов $X = \{x_1, \dots, x_n\}$, данные о которых представлены в виде матрицы «объект-объект» попарного сходства элементов исследуемой совокупности, которое, в свою очередь, минимизирует целевую функцию $Q(\tilde{T}) = \sum_{x_i, x_j \in X} (\mu_{\tilde{T}}(x_i, x_j) - \mu_T(x_i, x_j))^2$, так что нечеткую иерархию на множестве $X = \{x_1, \dots, x_n\}$ образуют четкие множества, полученные при декомпозиции отыскиваемого нечеткого отношения $\tilde{T}$. Особенностью процедуры, имеющей эвристический характер, является то,

что в зависимости от порядка α -уровней матрицы нечеткой толерантности T , представляющей исходные данные, возможна как восходящая, так и нисходящая версия алгоритма.

Предложенный Д. Думитреску FDN алгоритм [81] представляет собой бинарную дивизимную кластер-процедуру, основанную на обобщении FCM алгоритма.

Нечеткие иерархические кластер-процедуры обладают общей для всех иерархических процедур особенностью: с ростом количества объектов классифицируемой совокупности быстро растут время вычислений и требования к объему оперативной памяти ЭВМ, так что эти алгоритмы применимы для классификации совокупностей объектов сравнительно небольшого объема. Вместе с тем, их использование позволяет проводить более детальное исследование структуры классифицируемой совокупности.

3.3.2. Описание алгоритмов

Как и при описании нечетких эвристических кластер-процедур, при рассмотрении нечетких иерархических методов автоматической классификации перед изложением схемы алгоритма целесообразно рассматривать основные понятия и определения соответствующего иерархического метода.

3.3.2.1. Алгоритм Ватады — Танаки — Асаи [182] является эвристической процедурой и строит иерархию четких кластеров по уровням матрицы транзитивного отношения, вычисляемой на основе матрицы исходных данных, в свою очередь, представляющей собой некоторую нечеткую толерантность.

Основные понятия и определения

Пусть $X = \{x_1, \dots, x_n\}$ — исследуемая совокупность объектов, данные о которых представлены в виде матрицы «объект-объект» (1.3), элементы которой r_{ij} представляют собой значения нечеткой толерантности T , определенной на множестве X , то есть $r_{ij} = [\mu_T(x_i, x_j)]$, $i, j = 1, \dots, n$. Если элементы $\mu_T(x_i, x_j)$, $i, j = 1, \dots, n$ матрицы нечеткого отношения T упорядочить в возрастающем порядке, так что

$$\mu_T(x_1, x_1) < \mu_T(x_2, x_2) < \dots < \mu_T(x_k, x_k) < \dots, \quad (3.97)$$

где $\mu_T(x_k, x_k)$ располагается в k -й позиции в последовательности (3.97), то для обозначения номера этой позиции будет использоваться обозначение $\#(\mu_T(x_k, x_k))$, так что $\#(\mu_T(x_k, x_k)) = k$.

Задача заключается в построении такого нечеткого отношения $\tilde{T}$ на множестве, которое минимизирует функционал

$$Q(\tilde{T}) = \sum_{x_i, x_j \in X} (\mu_{\tilde{T}}(x_i, x_j) - \mu_T(x_i, x_j))^2, \quad (3.98)$$

при выполнении условия (max-min)-транзитивности (2.26). Задача минимизации критерия $Q(\tilde{T})$ (3.98) представляет собой задачу нелинейного программирования, так что при указанном ограничении вполне естественно для минимизации $Q(\tilde{T})$ использовать эвристический метод. С этой целью авторы метода вводят понятия max-транзитивной матрицы $\tilde{T}$, которая представляет собой матрицу (max-min)-транзитивного замыкания нечеткого отношения T и min-транзитивной матрицы $\hat{T}$, которая представляет собой матрицу, двойственную матрице $\tilde{T}$. Иными словами, матрица $\tilde{T}$ строится рекурсивно путем увеличения значения элементов $\mu_T(x_i, x_k)$, что авторы метода определяют следующим образом:

$$\begin{aligned} \mu_T(x_i, x_k) &< \bigvee_{x_j \in X} \mu_T(x_i, x_j) \wedge \mu_T(x_j, x_k) \Rightarrow \\ &\Rightarrow \mu_T(x_i, x_k) \geq \bigvee_{x_j \in X} \mu_T(x_i, x_j) \wedge \mu_T(x_j, x_k), \end{aligned} \quad (3.99)$$

а матрица $\hat{T}$ строится рекурсивно путем уменьшения значения элементов $\mu_T(x_i, x_j)$ и/или $\mu_T(x_j, x_k)$, так что

$$\begin{aligned} \mu_T(x_i, x_k) &< \mu_T(x_i, x_j) \wedge \mu_T(x_j, x_k) \Rightarrow \\ &\Rightarrow \mu_T(x_i, x_k) \geq \mu_T(x_i, x_j) \wedge \mu_T(x_j, x_k). \end{aligned} \quad (3.100)$$

В качестве примера [182, с.152-153] рассмотрим матрицу нечеткого отношения F и соответствующих ей матриц $\tilde{F}$ и $\hat{F}$:

$$F = \begin{pmatrix} 0.9 & 0.2 & 0.6 & 0.5 \\ 0.3 & 0.3 & 0.4 & 0.4 \\ 0.5 & 0.8 & 0.2 & 0.1 \\ 0.4 & 0.7 & 0.1 & 0.8 \end{pmatrix}, \quad \tilde{F} = \begin{pmatrix} 0.9 & 0.6 & 0.6 & 0.5 \\ 0.4 & 0.4 & 0.4 & 0.4 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.4 & 0.7 & 0.4 & 0.8 \end{pmatrix}, \quad \hat{F} = \begin{pmatrix} 0.9 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.8 \end{pmatrix}.$$

Таким образом, искомая матрица $\tilde{T}$, минимизирующая $Q(\tilde{T})$, отыскивается рекурсивно путем увеличения и/или уменьшения значений элементов исходной матрицы T , так что

$$\hat{T} \leq \tilde{T} \leq \tilde{T}. \quad (3.101)$$

Процедура построения матрицы $\tilde{T}$ предусматривает следующие основные этапы:

- 1 Вычисляются α -уровни, $\alpha \in [0,1]$, нечеткой толерантности T и для каждого $\alpha, \alpha \in [0,1]$ строится матрица T_α в соответствии с формулой (2.38);
- 2 Вводятся три индекса, относящиеся к измерению «совершенствования» транзитивности, и выбирается одна из двух стратегий декомпозиции;
- 3 Для каждого $\alpha, \alpha \in [0,1]$ вычисляется транзитивная матрица $\tilde{T}_\alpha$, соответствующая трем введенным индексам;
- 4 Строится транзитивная матрица $\tilde{T}$ по всем α -уровням $\tilde{T}_\alpha$ в соответствии с формулой (2.39).

α -уровень нечеткой толерантности T представляет собой не что иное, как уровень декомпозиции отношения T , так что в соответствии с теоремой декомпозиции отношения T_α для $\alpha_1 < \dots < \alpha_e < \dots < \alpha_Z$ образуют иерархию H ; аналогично соответствующие иерархии образуют отношения $\hat{T}_\alpha$, $\tilde{T}_\alpha$ и $\check{T}_\alpha$. Таким образом, стратификационный индекс иерархии H транзитивных матриц $\tilde{T}_\alpha$ для любого $\alpha_e, e = 1, \dots, Z$ в последовательности $\alpha_1 < \dots < \alpha_e < \dots < \alpha_Z$ может быть определен как $\ell(\alpha_e) = e$, так что упомянутые выше две стратегии декомпозиции могут быть определены следующим образом:

- 1) строятся $\tilde{T}_\alpha$ для $\ell(\alpha) = 1$ и $\tilde{T}_\alpha$ для $\ell(\alpha) = 2$ такие, что матрица $\alpha_1 \cdot \tilde{T}_{\alpha_1} \vee \alpha_2 \cdot \tilde{T}_{\alpha_2}$ является транзитивной, после чего отыскивается $\tilde{T}_\alpha$ для $\ell(\alpha) = 3$ такая, что матрица $\alpha_1 \cdot \tilde{T}_{\alpha_1} \vee \alpha_2 \cdot \tilde{T}_{\alpha_2} \vee \alpha_3 \cdot \tilde{T}_{\alpha_3}$ также транзитивна, и так далее вплоть до $\ell(\alpha) = Z$; данная стратегия именуется восходящим анализом;
- 2) строятся $\tilde{T}_\alpha$ для $\ell(\alpha) = Z$ и $\tilde{T}_\alpha$ для $\ell(\alpha) = Z - 1$ такие, что матрица $\alpha_Z \cdot \tilde{T}_{\alpha_Z} \vee \alpha_{Z-1} \cdot \tilde{T}_{\alpha_{Z-1}}$ является транзитивной, после чего отыскивается $\tilde{T}_\alpha$ для $\ell(\alpha) = Z - 2$ такая, что матрица $\alpha_Z \cdot \tilde{T}_{\alpha_Z} \vee \alpha_{Z-1} \cdot \tilde{T}_{\alpha_{Z-1}} \vee \alpha_{Z-2} \cdot \tilde{T}_{\alpha_{Z-2}}$ также транзитивна, и так далее вплоть до $\ell(\alpha) = 1$; данная стратегия именуется нисходящим анализом.

Процедура, реализующая восходящую стратегию, представляет собой дивизимную кластер-процедуру, тогда как процедура, реализующая нисходящую стратегию, представляет собой агломеративную

кластер-процедуру. Описанная ниже версия алгоритма реализует восходящую стратегию.

Очевидно, что для $\alpha_{e-1} \leq \alpha_e \leq \alpha_{e+1}$ выполняется

$$T_{\alpha_{e-1}} \geq T_{\alpha_e} \geq T_{\alpha_{e+1}}. \quad (3.102)$$

Матрицы $\hat{T}_\alpha$, $\tilde{T}_\alpha$ и $\check{T}_\alpha$ представляют собой α -уровни матриц $\hat{T}$, $\tilde{T}$ и $\check{T}$ соответственно, так что выполняется

$$\tilde{T}_\alpha \geq \check{T}_\alpha \geq \hat{T}_\alpha, \quad (3.103)$$

откуда можно выдвинуть следующее предложение: при $\alpha_{e-1} < \alpha_e < \alpha_{e+1}$ имеет место

$$\mu_{\hat{T}_{\alpha_e}}(x_i, x_j) \leq \mu_{\tilde{T}_{\alpha_e}}(x_i, x_j) \leq \mu_{\check{T}_{\alpha_{e-1}}}(x_i, x_j) \wedge \mu_{\tilde{T}_{\alpha_e}}(x_i, x_j), \quad (3.104)$$

$e=1, \dots, Z, \forall x_i, x_j \in X$

$$\mu_{\hat{T}_{\alpha_e}}(x_i, x_j) \vee \mu_{\tilde{T}_{\alpha_{e+1}}}(x_i, x_j) \leq \mu_{\check{T}_{\alpha_e}}(x_i, x_j) \leq \mu_{\tilde{T}_{\alpha_e}}(x_i, x_j), \quad (3.105)$$

$e=1, \dots, Z, \forall x_i, x_j \in X$

Пусть для рассмотренной выше матрицы F при некотором $\ell(\alpha) = e$ матрицы $\tilde{F}_{\alpha_e}$, $\hat{F}_{\alpha_e}$ и $\check{F}_{\alpha_{e-1}}$ определяются следующим образом [182, с.155]:

$$\tilde{F}_{\alpha_e} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \hat{F}_{\alpha_e} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \check{F}_{\alpha_{e-1}} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

тогда, учитывая, что

$$\tilde{F}_{\alpha_{e-1}} \wedge \check{F}_{\alpha_e} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

так что, в соответствии с (3.104), имеет место неравенство $\hat{F}_{\alpha_e} \leq \tilde{F}_{\alpha_e} \leq \check{F}_{\alpha_{e-1}} \wedge \check{F}_{\alpha_e}$, откуда

$$\tilde{F}_{\alpha_e} = \begin{pmatrix} 1 & * & 0 & 0 \\ * & * & 0 & 0 \\ 1 & 1 & * & * \\ * & 1 & * & 1 \end{pmatrix},$$

то есть задача заключается в нахождении и подстановке в матрице $\tilde{F}_{\alpha_e}$ вместо символов * таких значений $\mu_{\tilde{F}_{\alpha_e}}(x_i, x_j) \in \{0,1\}, i, j \in \{1,2,3,4\}$, чтобы целевая функция $Q(\tilde{T})$ достигала минимума.

α -уровень $T_\alpha, \alpha \in [0,1]$ нечеткой толерантности T представляет собой матрицу, для элементов $(x_i, x_k), (x_i, x_j), (x_j, x_k), i, j, k = 1, \dots, n$ которой справедливо $\mu_{T_\alpha}(x_i, x_k) = 0, \mu_{T_\alpha}(x_i, x_j) = 1, \mu_{T_\alpha}(x_j, x_k) = 1$. Нетранзитивность T_α также может быть выражена условием

$$\mu_{T_\alpha}(x_i, x_k) < \mu_{T_\alpha}(x_i, x_j) \wedge \mu_{T_\alpha}(x_j, x_k). \quad (3.106)$$

Для приведения матрицы T_α к транзитивной матрице авторами предлагаются следующие преобразования:

- 1) *изменение 1*: в левой части выражения (3.106) элемент $\mu_{T_\alpha}(x_i, x_j)$ изменяется с 0 на 1;
- 2) *изменение 2*: в правой части выражения (3.106) изменяется с 1 на 0 элемент $\mu_{T_\alpha}(x_i, x_k)$ и/или элемент $\mu_{T_\alpha}(x_k, x_j)$;

Для получения max-транзитивной матрицы $\tilde{T}_\alpha$ используется изменение 1, а для получения min-транзитивной матрицы $\hat{T}_\alpha$ используется изменение 2; в дальнейшем будет показано, что для получения матрицы $\tilde{T}_\alpha$ используются оба изменения.

Для выявления элементов матрицы T_α , подлежащих изменению, вводятся три индекса. Помимо обнаружения изменяемых элементов матрицы T_α , эти индексы определяют вид применяемого изменения.

Индекс 1 уровня декомпозиции $\ell(\alpha)$ имеет обозначение $I_1^\ell(x_i, x_j)$ и определяется выражением

$$I_1^\ell(x_i, x_j) = \begin{cases} \sum_{k=1}^n \mu_{T_\alpha}(x_i, x_k) \wedge \mu_{T_\alpha}(x_k, x_j), & \text{если } \mu_{T_\alpha}(x_i, x_j) = 0 \\ 0, & \text{если } \mu_{T_\alpha}(x_i, x_j) = 1 \end{cases}. \quad (3.107)$$

$I_1^\ell(x_i, x_j)$ показывает те элементы, к которым необходимо применить изменение 1: в соответствии с изменением 1 должны быть преобразованы те элементы (x_i, x_j) матрицы T_α , для которых значение $I_1^\ell(x_i, x_j)$ является максимальным.

Например, если при некотором $\ell(\alpha) = e$ для рассмотренной выше матрицы F матрица F_{α_e} будет иметь вид

$$F_{\alpha_e} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix},$$

то матрица $I_1^t(x_i, x_j)$ для F_{α_e} будет выглядеть следующим образом:

$$I_1^t(x_i, x_j) = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Индекс 2 уровня декомпозиции $\ell(\alpha)$ имеет обозначение $I_2^t(x_i, x_j)$ и определяется выражением

$$I_2^t(x_i, x_j) = \sum_{k=1, I_1^t(x_i, x_k) \neq 0}^n \frac{\mu_{T_\alpha}(x_i, x_j) \wedge \mu_{T_\alpha}(x_j, x_k)}{M^{I_1^t(x_i, x_k)}} + \\ + \sum_{k=1, I_1^t(x_k, x_j) \neq 0}^n \frac{\mu_{T_\alpha}(x_k, x_i) \wedge \mu_{T_\alpha}(x_i, x_j)}{M^{I_1^t(x_k, x_j)}}. \quad (3.108)$$

где M в знаменателе каждого из слагаемых представляет собой некоторое достаточно большое число. Так же, как $I_1^t(x_i, x_j)$ показывает эффективность применения изменения 1 к элементам матрицы T_α , индекс $I_1^t(x_i, x_j)$ показывает эффективность применения к ним изменения 2.

Индекс 3 уровня декомпозиции $\ell(\alpha)$ имеет обозначение $I_3^t(x_i, x_j)$ и определяется в зависимости от стратегии декомпозиции. Если декомпозиция производится в соответствии со стратегией восходящего анализа, то $I_3^t(x_i, x_j)$ определяется в соответствии с формулой

$$I_3^t(x_i, x_j) = \begin{cases} \#(\mu_T(x_i, x_j)) + 1; \mu_{\tilde{T}_{\alpha_e}}(x_i, x_j) < \mu_{\tilde{T}_{\alpha_{e-1}}}(x_i, x_j) \wedge \mu_{\tilde{T}_{\alpha_e}}(x_i, x_j) \\ 0; \text{иначе} \end{cases}. \quad (3.109)$$

Если же декомпозиция производится в соответствии со стратегией нисходящего анализа, то $I_3^t(x_i, x_j)$ определяется в соответствии с формулой

$$I_3^t(x_i, x_j) = \begin{cases} \#(\mu_T(x_i, x_j)) + 1; \mu_{T_{\alpha_e}}(x_i, x_j) \vee \mu_{T_{\alpha_{e+1}}}(x_i, x_j) < \mu_{T_{\alpha_e}}(x_i, x_j) \\ 0; \text{иначе} \end{cases}. \quad (3.110)$$

В случае $I_3'(x_i, x_j) = 0$, так что $\mu_{\tilde{f}_{\alpha_e}}(x_i, x_j) = \mu_{\tilde{f}_{\alpha_e-1}}(x_i, x_j) \wedge \mu_{\tilde{f}_{\alpha_e}}(x_i, x_j)$, должно выполняться равенство $\mu_{\tilde{f}_{\alpha_e}}(x_i, x_j) = \mu_{\tilde{f}_{\alpha_e}}(x_i, x_j)$. В свою очередь, в случае, когда $\mu_{\tilde{f}_{\alpha_e}}(x_i, x_j) < \mu_{\tilde{f}_{\alpha_e-1}}(x_i, x_j) \wedge \mu_{\tilde{f}_{\alpha_e}}(x_i, x_j)$, значение индекса 3, соответствующее элементам, обозначенным символом * в матрице $\tilde{F}_{\alpha_e}$ в рассмотренном выше примере, принимает значение $\#(\mu_T(x_i, x_j)) + 1$.

В случае, когда $I_3'(x_i, x_j) \leq \ell(\alpha)$, то $\mu_{T_\alpha}(x_i, x_j) = 0$, а когда $I_3'(x_i, x_j) > \ell(\alpha)$, то $\mu_{T_\alpha}(x_i, x_j) = 1$. Таким образом, изменение 1 должно быть применено к элементу (x_i, x_j) матрицы T_α , минимизируя выражение $|\ell(\alpha) - I_3'(x_i, x_j)|$ при условии выполнения $0 < I_3'(x_i, x_j) < \ell(\alpha)$, тогда как изменение 2 должно быть применено к элементу (x_i, x_j) матрицы T_α , минимизируя выражение $|\ell(\alpha) - I_3'(x_i, x_j)|$ при условии, что выполняется неравенство $I_3'(x_i, x_j) > \ell(\alpha)$.

Например, для рассмотренных выше матрицы F и матриц $\tilde{F}_{\alpha_e}$, $\hat{F}_{\alpha_e}$ и $\tilde{F}_{\alpha_e-1}$, определенных для F при некотором $\ell(\alpha) = e$, матрица $I_3'(x_i, x_j)$ будет выглядеть следующим образом:

$$I_3'(x_i, x_j) = \begin{pmatrix} 0 & 3 & 0 & 0 \\ 4 & 4 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 5 & 0 & 2 & 0 \end{pmatrix}.$$

Индексы $I_1'(x_i, x_j)$ и $I_2'(x_i, x_j)$ обозначают различные аспекты нетранзитивности элементов, и взаимосвязь индексов $I_1'(x_i, x_j)$ и $I_2'(x_i, x_j)$ выражается соотношением

$$\sum_{i=1}^n \sum_{j=1}^n I_1'(x_i, x_j) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sum_{h=1}^n A_h(x_i, x_j), \quad (3.111)$$

где $A_h(x_i, x_j) = [I_1'(x_i, x_j) \cdot M^h] - [I_2'(x_i, x_j) \cdot M^{h-1}] \cdot M$. Справедливость соотношения (3.111) следует из преобразования

$$\begin{aligned}
A_h(x_i, x_j) &= [I_1^\ell(x_i, x_j) \cdot M^h] - [I_2^\ell(x_i, x_j) \cdot M^{h-1}] \cdot M = \\
&= \sum_{k=1, I_1^\ell(x_i, x_k)=h}^n \mu_{T_\alpha}(x_i, x_j) \wedge \mu_{T_\alpha}(x_j, x_k) + \\
&+ \sum_{k=1, I_1^\ell(x_k, x_j)=h}^n \mu_{T_\alpha}(x_k, x_i) \wedge \mu_{T_\alpha}(x_i, x_j)
\end{aligned}$$

а также

$$\begin{aligned}
\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n A_h(x_i, x_j) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1, I_1^\ell(x_i, x_k) > 0}^n \mu_{T_\alpha}(x_i, x_j) \wedge \mu_{T_\alpha}(x_j, x_k) + \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1, I_1^\ell(x_k, x_j) > 0}^n \mu_{T_\alpha}(x_k, x_i) \wedge \mu_{T_\alpha}(x_i, x_j) = \\
&= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^n I_1^\ell(x_i, x_k) + \frac{1}{2} \sum_{k=1}^n \sum_{j=1}^n I_1^\ell(x_k, x_j) = \sum_{i=1}^n \sum_{j=1}^n I_1^\ell(x_i, x_j)
\end{aligned}$$

Таким образом, выражение (3.111) указывает, что сумма значений $I_1^\ell(x_i, x_j)$, каждое из которых показывает эффективность применения изменения 1, эквивалентна сумме значений $I_2^\ell(x_i, x_j)$, каждое из которых, в свою очередь, показывает эффективность применения изменения 2.

Параметры алгоритма:

Исходные данные задаются в виде матрицы нечеткой толерантности $r_{n \times n} = [\mu_T(x_i, x_j)]$, $i, j = 1, \dots, n$. Входные параметры как таковые отсутствуют.

Схема алгоритма:

- 1 Для исходной матрицы $r_{n \times n} = [\mu_T(x_i, x_j)]$ строится матрица (max-min)-транзитивного замыкания $\tilde{T}$ и min-транзитивная матрица $\hat{T}$ в соответствии с формулами (3.99) и (3.100) соответственно;
- 2 Вычисляются значения α -уровней, $\alpha_e, e = 1, \dots, Z$, и производится их нумерация, так что $\ell(\alpha_e) = e$; выбирается порог α такой, что $\ell(\alpha) = 1$, и полагается $N := 1$;
- 3 Для уровня $\ell(\alpha) = e$ производится декомпозиция нечетких матриц $T, \tilde{T}, \hat{T}$ для получения матриц $T_{\alpha_e}, \tilde{T}_{\alpha_e}, \hat{T}_{\alpha_e}$ соответственно;
- 4 Для матриц $\tilde{T}_{\alpha_{e-1}}, \tilde{T}_{\alpha_e}$ и $\hat{T}_{\alpha_e}$ строится $I_3^\ell(x_i, x_j)$ в соответствии с формулой (3.109);

- 5 Для матрицы T_{α_e} строятся $I_1^t(x_i, x_j)$ и $I_2^t(x_i, x_j)$ в соответствии с формулами (3.107) и (3.108);
- 6 Выбирается элемент (x_i, x_j) , которому в матрицах $I_1^t(x_i, x_j)$ и $I_2^t(x_i, x_j)$ соответствует наибольшее положительное число;
- 7 Если на предыдущем шаге выбран по крайней мере один элемент (x_i, x_j) , то осуществляется переход на шаг 8, в противном случае осуществляется переход на шаг 10;
- 8 К элементам, выбранным на шаге 6, применяются изменение 1 и/или изменение 2;
- 9 Если матрица T_{α_e} транзитивна, осуществляется переход на шаг 15, в противном случае осуществляется переход на шаг 5;
- 10 Если для $I_3^t(x_i, x_j)$ превышено предельное значение $\#(\mu_T(x_k, x_k))$ в последовательности, определяемой формулой (3.97), осуществляется переход на шаг 12, в противном случае осуществляется переход на шаг 11;
- 11 Полагается $N := N + 1$, где N обозначает откорректированное предельное значение $\#(\mu_T(x_k, x_k))$ индекса $I_3^t(x_i, x_j)$, так что $|I_3^t(i, j) - \ell(\alpha_e)| \leq N$, и осуществляется переход на шаг 6;
- 12 Если число итераций для реконструкции $I_3^t(x_i, x_j)$ превышает заданное, осуществляется переход на шаг 14, в противном случае осуществляется переход на шаг 13;
- 13 Производится реконструкция $I_3^t(x_i, x_j)$, так что измененные элементы просматриваются опять, полагается $N := 1$, число итераций увеличивается на 1 и осуществляется переход на шаг 11;
- 14 Полагается $\tilde{T}_{\alpha_e} := \hat{T}_{\alpha_e}$ и осуществляется переход на шаг 16;
- 15 Полагается $\tilde{T}_{\alpha_e} := T_{\alpha_e}$;
- 16 Составляется матрица $\tilde{T}$ в соответствии с правилом $\tilde{T} = \bigvee_{\alpha_e} (\alpha_e \cdot \tilde{T}_{\alpha_e})$;
- 17 Если $\ell(\alpha) > Z$, осуществляется переход на шаг 19, в противном случае осуществляется переход на шаг 18;
- 18 Полагается $\ell(\alpha) := \ell(\alpha) + 1$ и осуществляется переход на шаг 3;
- 19 Алгоритм прекращает работу.

В работе [182] приводится пример применения процедуры к анализу совокупности пятнадцати различных прохладительных напитков. В сравнении с иерархической версией процедуры Тамуры — Хигути — Танаки алгоритм показал гораздо лучшие результаты.

3.3.2.2. Алгоритм Думитреску (FDH algorithm) [81] является дивизимной процедурой, последовательно строящей бинарное дерево нечеткой иерархии.

Основные понятия и определения

Пусть, как и ранее, $X = \{x_1, \dots, x_n\}$ — исследуемая совокупность объектов, данные о которых представлены в виде матрицы «объект-свойство» $X_{m \times n} = [x'_i], i = 1, \dots, n, t = 1, \dots, m$, то есть каждый объект представляет собой точку в пространстве признаков $I^m(X)$. Кластерная структура исследуемого множества объектов X описывается семейством s разделенных нечетких множеств, образующих разбиение $P = \{A^1, \dots, A^c\}$ на нечеткие кластеры. Если некоторый нечеткий кластер $A^i \in P$ неоднороден, то можно рассматривать его субструктуру, которая, в свою очередь, также может быть описана некоторым разбиением нечеткого множества A^i , так что образуется многоуровневая нечеткая классификация, и каждое последующее разбиение является уточнением разбиения, соответствующего предыдущему уровню классификации. Перед рассмотрением основных определений, введенных Д. Думитреску, следует оговорить, что операции над нечеткими множествами, в частности, рассмотренные в главе 2 операции пересечения и объединения нечетких множеств, в общем, определяются с помощью треугольных норм (t-норм), обозначаемых символом T , и треугольных конорм (t-конорм, s-норм), обозначаемых символом S [34, с.29-36]; однако в данном случае полное рассмотрение треугольных норм нецелесообразно. При разработке процедуры Д. Думитреску [81, с.146] определил операции пересечения и объединения нечетких множеств с помощью треугольных норм и треугольных конорм следующим образом:

- 1) Пересечение нечетких множеств A и B определяется в соответствии с формулой

$$\mu_{A \cap B}(x) = T(\mu_A(x), \mu_B(x)), \forall x \in X, \quad (3.112)$$

где $T(\mu_A(x), \mu_B(x)) = \max(0, \mu_A(x) + \mu_B(x) - 1), \forall x \in X$.

- 2) Объединение нечетких множеств A и B определяется в соответствии с формулой

$$\mu_{A \cup B}(x) = S(\mu_A(x), \mu_B(x)), \forall x \in X, \quad (3.113)$$

где $S(\mu_A(x), \mu_B(x)) = \min(1, \mu_A(x) + \mu_B(x)), \forall x \in X$.

При дальнейшем рассмотрении метода, предложенного Д. Думитреску, будет приниматься, что операции пересечения и объединения нечетких множеств определены с помощью (3.112) и (3.113).

Нечеткие множества $A^1, A^2, \dots, A^c, c \geq 2$ называются разделенными, если для $k = 1, \dots, c-1$ выполняется условие $\left(\bigcup_{i=1}^k A^i\right) \cap A^{k+1} = \emptyset$.

Семейство разделенных нечетких множеств $P^i = \{A^1, \dots, A^c\}$ образует разбиение некоторого нечеткого множества A^i , если выполняется условие $A^i = \bigcup_{k=1}^c A_k^i$, так что при $A^i = X$ приведенное условие эквивалентно условию нечеткого разбиения, введенного Э. Распини,

$$\sum_{i=1}^c \mu_{A^i} = 1, \text{ выполняющегося для каждого } x_i \in X.$$

Поскольку символом Π обозначается множество всех нечетких разбиений исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, то множество всех нечетких разбиений $P_b^i, b = 1, \dots, f$ некоторого нечеткого множества A^i будет обозначаться символом Π^i и пусть имеет место $P_1^i, P_2^i \in \Pi^i, P_1^i = \{A_{(1)1}^i, \dots, A_{(u)1}^i\}, P_2^i = \{A_{(1)2}^i, \dots, A_{(z)2}^i\}$. Нечеткое разбиение P_2^i предшествует нечеткому разбиению P_1^i , или, как выражается автор процедуры [81, с.146], «уточняет» нечеткое разбиение P_1^i , что обозначается $P_1^i \prec P_2^i$, в том случае, если каждый элемент, или, по выражению автора алгоритма, атом разбиения P_1^i может быть представлен объединением атомов разбиения P_2^i , что можно записать также следующим образом: $P_1^i \prec P_2^i$, если и только если $z \geq u$ и существует разбиение $\{K(k) | k = 1, 2, \dots, u\}$ множества $\{1, 2, \dots, z\}$, такое, что выполняется условие $A_{(i)1}^i = \bigcup_{k \in K(k)} A_{(k)2}^i, i \in K(k)$ и условие

$$A_{(k)1}^i \cap \left(\bigcup_{k \in K(k)} A_{(k)2}^i \right) = \emptyset, i = 1, \dots, u.$$

Если $P = \{A^1, \dots, A^c\}$ является нечетким разбиением исследуемой совокупности объектов X , то можно допустить, что каждый не-

четкий кластер $A^l, l=1, \dots, c$ может быть представлен прототипом, или, иными словами, центром $\tau^l, l=1, \dots, c$.

Пусть $d(x_i, \tau^l)$ — функция расстояния, измеряющая степень отличия точки x_i от l -го прототипа τ^l и представляющая собой квадрат евклидова расстояния: $d(x_i, \tau^l) = \|x_i - \tau^l\|^2$. При обозначении собственно евклидова расстояния между объектами x_i и x_j через $D(x_i, x_j)$ и определении понятия локального расстояния $D_l(x_i, x_j)$ в пространстве признаков $I^m(X)$, зависящего от функции принадлежности нечеткого множества A^l , соотношением

$$D_l(x_i, x_j) = \begin{cases} \min(\mu_{ii}, \mu_{jj}) D(x_i, x_j), & x_i, x_j \in X \\ \mu_{ii} D(x_i, x_j), & x_i \in X, x_j \notin X \\ D(x_i, x_j), & x_i, x_j \notin X \end{cases} \quad (3.114)$$

локальная функция расстояния $d_l(x_i, x_j)$ будет определяться, в свою очередь, соотношением

$$d_l(x_i, x_j) = D_l^2(x_i, x_j). \quad (3.115)$$

В таком случае локальное различие между некоторым объектом $x_i \in X$ и центром τ^l нечеткого кластера A^l будет определяться выражением

$$d_l(x_i, \tau^l) = \mu_{ii}^2 d(x_i, \tau^l) = \mu_{ii}^2 \|x_i - \tau^l\|^2, \quad (3.116)$$

а мера неадекватности между нечетким кластером A^l и его центром τ^l будет определяться выражением

$$I(A^l, \tau^l) = \sum_{i=1}^n \mu_{ii}^2 d(x_i, \tau^l) = \sum_{i=1}^n \mu_{ii}^2 \|x_i - \tau^l\|^2, \quad (3.117)$$

так что мера неадекватности между разбиением $P = \{A^1, \dots, A^c\}$ и набором его центров $\tau^1, \dots, \tau^c$ будет определяться формулой

$$Q_D^n(P) = \sum_{l=1}^c I(A^l, \tau^l) = \sum_{l=1}^c \sum_{i=1}^n \mu_{ii}^2 d(x_i, \tau^l) = \sum_{l=1}^c \sum_{i=1}^n \mu_{ii}^2 \|x_i - \tau^l\|^2, \quad (3.118)$$

представляющей собой ни что иное, как функционал $Q_1^n(P)$ Дж. Беж-дека — Дж. Данна в виде (3.46) при $\gamma = 2$, так что, естественно, интегрес представляет задача минимизации функционала (3.118).

Кластерная структура некоторого нечеткого класса $A^l \in P, P \in \Pi$ может быть описана его нечетким разбиением. Разбиение неоднородных нечетких кластеров — элементов нечеткого разбиения P приво-

дит к новому нечеткому разбиению P^l , «уточняющему» P . Оптимальное нечеткое разбиение $P^l = \{A_1^l, \dots, A_c^l\}$ с функциями принадлежности $\mu_{k_i}^{(l)}, k=1, \dots, c$ нечеткого класса A^l может быть найдено как решение оптимизационной задачи $Q_D^H(P^l) \rightarrow \min_{P^l \in \Pi_c^l}$, где Π_c^l — семейство всех нечетких разбиений нечеткого кластера A^l на c классов; нечеткое разбиение $P^l \in \Pi_c^l$ будет доставлять минимум целевой функции, которую теперь можно переписать в виде

$$Q_D^H(P^l) = \sum_{k=1}^c \sum_{i=1}^n (\mu_{k_i}^{(l)})^2 \|x_i - \tau^{k(l)}\|^2, \quad (3.119)$$

если выполняются условия

$$\tau^{k(l)} = \sum_{i=1}^n (\mu_{k_i}^{(l)})^2 \cdot x_i / \sum_{i=1}^n (\mu_{k_i}^{(l)})^2, k=1, \dots, c, \quad (3.120)$$

$$\mu_{k_i}^{(l)} = \frac{\mu_{k_i}}{\sum_{a=1}^c \left(\frac{\|x_i - \tau^{k(l)}\|}{\|x_i - \tau^{a(l)}\|} \right)^2}, k=1, \dots, c, i=1, \dots, n, \quad (3.121)$$

где $\tau^{k(l)}$ — центр нечеткого кластера $A_k^l \subseteq A^l$, $\mu_{k_i}^{(l)}$ — степень принадлежности некоторой точки $x_i \in X$ нечеткому кластеру A_k^l , а μ_{k_i} — степень принадлежности некоторой точки $x_i \in X$ нечеткому кластеру A^l .

Таким образом, используя в алгоритме Беждека — Данна соотношения (3.120) и (3.121) для вычисления центров нечетких кластеров и построения нечетких разбиений, можно получить процедуру, которую Д. Думитреску назвал обобщенной нечеткой процедурой ISO-DATA (Generalized Fuzzy ISODATA), или, сокращенно, GFI алгоритмом [81, с.148]. В частности, при $A^l = X$ данная процедура превращается в описанный выше алгоритм Беждека — Данна при $\gamma = 2$.

Далее, пусть при помощи GFI — алгоритма получено разбиение некоторого нечеткого кластера A^l на два кластера, так что $P^l = \{A_1^l, A_2^l\}$. Если A^l является компактным кластером, то для всех точек степени принадлежности обоим атомам разбиения $P^l = \{A_1^l, A_2^l\}$ оказываются приблизительно одинаковы, и наоборот, если эти атомы представляют собой хорошо разделенные кластеры, то значения принадлежностей $\mu_{1i}^{(l)}$ и $\mu_{2i}^{(l)}$ поляризуются у значений 0 и μ_{k_i} для всех $x_i \in X$, так что структурированность данных подразумевает поляри-

зацию, и степень поляризации нечеткого разбиения может рассматриваться как мера качества разбиения.

Пусть $P_1^i, P_2^i \in \Pi_2^i$ — разбиения нечеткого кластера A^i на два кластера, так что $P_1^i = \{A_{1(1)}^i, A_{2(1)}^i\}$ и $P_2^i = \{A_{1(2)}^i, A_{2(2)}^i\}$, где $\mu_{ki}^{b(i)}$ — значение функции принадлежности i -й точки k -му нечеткому кластеру — атому b -го нечеткого разбиения P_b^i нечеткого кластера A^i . Нечеткое разбиение P_1^i поляризовано меньше, чем нечеткое разбиение P_2^i , что будет обозначаться $P_1^i \leq_p P_2^i$, в том и только в том случае, если для всех $x_i \in X$ выполняется условие

$$|\mu_{1i}^{1(i)} - \mu_{2i}^{1(i)}| \leq |\mu_{1i}^{2(i)} - \mu_{2i}^{2(i)}|, i = 1, \dots, n, \quad (3.122)$$

так что отношение $\leq_p$ представляет собой определенный на Π_2^i предпорядок.

Пусть A^i — нечеткое множество и $P^i = \{A_1^i, A_2^i\}$ — нечеткое разбиение A^i . Степень поляризации бинарного нечеткого разбиения нечеткого множества A^i представляет собой функцию $\aleph: \Pi_2^i \rightarrow R$, удовлетворяющую следующим условиям:

- 1) $a \leq \aleph(P^i) \leq 1, \forall P^i \in \Pi_2^i, a \in [0, 1]$;
- 2) $\aleph(P^i) = 1 \Leftrightarrow$ либо $\mu_{1i}^{(i)} = \mu_{ii} > 0.5$, либо $\mu_{2i}^{(i)} = \mu_{ii} > 0.5, \forall x_i \in X$;
- 3) $A_1^i = A_2^i \Rightarrow \aleph(P^i) = a$;
- 4) $P_1^i \leq_p P_2^i \Rightarrow \aleph(P_1^i) \leq \aleph(P_2^i)$.

Из условия 2) очевидно, что если P является разбиением исследуемой совокупности объектов X , то $\aleph(P) = 1$ только в том случае, если P является четким разбиением X . Данное обстоятельство вместе с условием 3) позволяют рассматривать $\aleph(P)$ как своего рода «меру неразмытости» [81, с.149] нечеткого разбиения P .

Степени принадлежности компактным и хорошо разделенным, или, используя авторскую терминологию, «реальным», кластерам, полученные последовательной бинарной декомпозицией данных, сохраняют поляризацию в областях 0 и 1. Значение степени принадлежности в области 0.5 говорит об отсутствии структурированности данных.

Подобные рассуждения непосредственно подводят к следующему способу определения степени поляризации:

$$\aleph(P^i) = \frac{\sum_{x_i \in P^i} (\mu_{1i}^{(i)} + \mu_{2i}^{(i)})}{\sum_{x_i \in P^i} x_i}, \quad (3.123)$$

где, в свою очередь, $\mu_{ki}^{(i)}$ — значение принадлежности i -й точки нечеткому множеству $A_{(0.5)k}^i$ уровня 0.5, определяемому в соответствии с выражением (2.4), что в данной ситуации также можно записать в следующем виде:

$$\mu_{ki}^{(i)} = \begin{cases} \mu_{ki}^{(i)}, & \text{если } \mu_{ki}^{(i)} > 0.5 \\ 0, & \text{иначе} \end{cases}. \quad (3.124)$$

Знаменатель в выражении (3.123) представляет собой просто количество точек в кластерах $A_k^i, k=1,2$ разбиения P^i . Таким образом, для степени поляризации, определяемой в соответствии с выражением (3.123), $\alpha=0$ в силу условий 1) — 4), и можно допустить, что нечеткое разбиение $P^i = \{A_1^i, A_2^i\}$ описывает «реальные» кластеры в том и только в том случае, если для каждого класса $A_k^i, k=1,2$ существует x_i , такой, что $\mu_{ki}^{(i)} > 0.5$ и $\aleph(P^i) \geq \kappa$. В таком случае $\kappa \in (0,1)$ представляет собой некоторый порог, выражающий уверенность в том, что кластеры являются «реальными». Для иллюстрации вышеизложенного целесообразно рассмотреть следующий пример [81, с.151].

Пусть $X = \{x_1, x_2, x_3, x_4, x_5\}$ — исследуемая совокупность объектов и нечеткое разбиение $P = \{A^1, A^2\}$ совокупности X описывается матрицей, представленной таблицей 3.2.

Таблица 3.2

Матрица нечеткого разбиения P совокупности X на два класса

Кластер	Объект				
	x_1	x_2	x_3	x_4	x_5
A^1	0.95	0.98	0.10	0.18	0.05
A^2	0.05	0.02	0.90	0.82	0.95

В соответствии с (3.123) $\aleph(P) = 4.6/5 = 0.92$, так что при $\kappa \leq 0.92$ нечеткие кластеры, описываемые разбиением P , могут рассматриваться как «реальные».

Далее, пусть $P^1 = \{A_1^1, A_2^1\}$ — нечеткое разбиение кластера A^1 , описываемое матрицей, представленной таблицей 3.3.

Матрица нечеткого разбиения P^1 кластера A^1 на два класса

Кластер	Объект				
	x_1	x_2	x_3	x_4	x_5
A_1^1	0.90	0.95	0.10	0.10	0.00
A_2^1	0.05	0.03	0.00	0.08	0.05

Несмотря на достаточно высокое значение степени поляризации $K(P^1) \approx 0.82$, значения принадлежностей всех объектов классу A_2^1 менее 0.5, так что разбиение P^1 не представляет «реальные» кластеры при любом значении порога k и нечеткий кластер A^1 является «реальным» неделимым кластером.

Теперь, пусть $P^2 = \{A_1^2, A_2^2\}$ — нечеткое разбиение кластера A^2 , описываемое матрицей, представленной таблицей 3.4.

Таблица 3.4

Матрица нечеткого разбиения P^2 кластера A^2 на два класса

Кластер	Объект				
	x_1	x_2	x_3	x_4	x_5
A_1^2	0.05	0.01	0.85	0.80	0.05
A_2^2	0.00	0.01	0.05	0.02	0.90

При вычислении значения степени поляризации $K(P^2) \approx 0.93$ становится очевидным, что A_1^2 и A_2^2 являются «реальными» кластерами при $k \leq 0.93$, причем нечеткий кластер A_2^2 содержит только объект x_5 .

Семейство нечетких множеств, определенных на исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, образует нечеткую иерархию H , если выполняются следующие условия:

- 1) $A^i \cap A^f = \emptyset$, либо $A^i \subseteq A^f$, $A^i \neq A^f$, либо $A^f \subseteq A^i$, $A^f \neq A^i$ для всех $A^i, A^f \in H$;
- 2) $\bigcup_{A_k^i \in S(A^i)} A_k^i = A^i$ для каждого $A^i \in H$, так что $S(A^i) \neq \emptyset$, где $A_k^i \in S(A^i)$

$$S(A^i) = \{A_k^i \in H \mid A_k^i \neq A^i, A_k^i \subseteq A^i,$$

$$\forall A^f \in H, A^f \neq X, A^f \neq A^i, A_k^i \subseteq A^f \Rightarrow A^f = A_k^i\}$$

Иными словами, $S(A^i)$ — множество всех нечетких подмножеств, составляющих нечеткое множество A^i . Нечеткая иерархия H

называется всюду определенной, если $X \in H$. Нечеткая иерархия H называется бинарной, если $\text{card}(S(A^i)) = 2$ или $\text{card}(S(A^i)) = 0$ для каждого $A^i \in H$; некоторый элемент A^i нечеткой иерархии H , для которого выполняется $\text{card}(S(A^i)) = 0$, называется минимальным по включению в H . Таким образом, всюду определенная бинарная нечеткая иерархия H может быть представлена бинарным деревом, имеющим X в качестве корневого узла, терминальные узлы которого, в свою очередь, будут соответствовать минимальным по включению в H множествам.

Стратификационный индекс ℓ всюду определенной нечеткой иерархии H представляет собой функцию $\ell: H \rightarrow R^+$, удовлетворяющую следующим условиям:

- 1) для любых $A^i, A^f \in H, A^i \subseteq A^f, A^i \neq A^f$ имеет место неравенство $\ell(A^i) > \ell(A^f)$;
- 2) $\ell(X) = 0$.

Стратификационный индекс, таким образом, представляет собой уровень декомпозиции исследуемой совокупности объектов X . Сущность подхода заключается в том, что рассматриваются только те нечеткие разбиения в вышеуказанном смысле, атомы которых являются «реальными» кластерами. Процедура начинается с вычисления нечеткого разбиения $P = \{A^1, A^2\}$ исследуемой совокупности объектов X , которую можно рассматривать как однородный кластер, центр которого вычисляется в соответствии с формулой

$$\tau_i = \sum_{i=1}^n x_i^t / n, t = 1, \dots, m, \quad (3.125)$$

где x_i^t представляет собой элемент матрицы «объект-свойство» (1.1), то есть значение t -го признака на i -м объекте.

Если атомы A^1 и A^2 не являются «реальными» кластерами, можно говорить об отсутствии кластерной структуры в исследуемой совокупности объектов X . Если же атомы A^1 и A^2 являются «реальными» кластерами, то полагается $P_{(1)}^1 = \{A_{(1)}^1, A_{(1)}^1\}$. Следует указать на небольшое изменение смысла обозначений: в обозначении разбиения $P_{(l)}^i$ индекс ℓ обозначает уровень декомпозиции, символ l — номер разбиения соответствующего уровня, а символ $A_{k(l)}^i$ обозначает k -й кластер, $k=1,2$, l -го разбиения ℓ -го уровня декомпозиции. Далее, используя GFI-алгоритм, вычисляем бинарные нечеткие разбиения

атомов $A_{k(l)}^1, k=1,2$ разбиения $P_{(1)}^1$: если атомы новых разбиений являются «реальными» кластерами, то они будут атомами разбиений уровня $\ell=2$, в противном случае атом, не являющийся «реальным» кластером, оказывается неделимым нечетким кластером уровня $\ell=2$. В свою очередь, атомы разбиения, не являющиеся неделимыми нечеткими кластерами, подвергаются дальнейшей обработке данной процедурой, напоминающей бэктрекинг в языке логического программирования PROLOG. Так, атомы предыдущего уровня декомпозиции будут именоваться формирующими нечеткие разбиения текущего уровня исследуемой совокупности объектов X . Процесс декомпозиции останавливается тогда, когда не обнаруживается ни одного «реального» кластера, то есть два следующих друг за другом разбиения оказываются идентичны: при $P_{(b)}^l = P_{(b-1)}^l$ для всех l уровень b оказывается последним уровнем декомпозиции X .

Последовательность разбиений различных уровней $\ell, 0,1,\dots,b$, полученных в результате декомпозиции, где разбиение уровня $\ell=0$ представляет собой исследуемую совокупность объектов X , что обозначается $P_{(0)} = \{X, \emptyset\}$, удовлетворяет рассмотренному выше условию предшествования, так что $P_{(\ell)} \prec P_{(\ell+1)}$, а всюду определенная бинарная нечеткая иерархия может быть представлена в виде $H = \{A_{k(\ell)}^i \mid A_{k(\ell)}^i \in P_{(\ell)}^i, \ell = \{0,1,\dots,b\}\}$. В свою очередь, стратификационный индекс определяется выражением

$$\ell(A_{k(\ell)}^i) = \min\{\ell \mid A \in P_{(\ell)}^i\}, \forall A_{k(\ell)}^i \in H. \quad (3.126)$$

Естественно, что число классов в разбиении уровня $\ell=0$ будет равно одному, а в разбиении уровня $\ell=1$ — двум, так что при $\ell=0$ и $\ell=1$ число собственно разбиений будет равно одному, следовательно, при обозначении разбиений $P_{(0)}^i$ и $P_{(1)}^i$ верхний индекс i можно опускать.

Параметры алгоритма:

Исходные данные задаются в виде матрицы «объект-свойство» (1.1). Входные параметры как таковые отсутствуют. В процессе работы алгоритма используется счетчик терминальных узлов c . Процесс присвоения номеров l разбиениям $P_{(\ell)}^i$ детально не рассматривается.

Схема алгоритма:

- 1 Полагается $\ell := 0; c = 0; P_{(0)} = \{X, \emptyset\}$;
- 2 Для текущего уровня ℓ нумеруются разбиения $l=1,\dots$; для каждого разбиения $P_{(\ell)}^l$ просматриваются все атомы, и для каждого не-

- пустого нечеткого кластера $A_{k(\ell)}^i \in P_{(\ell)}^i$ осуществляется выполнение условия, описанного на шаге 4; все кластеры $A_{k(\ell)}^i$ — терминальные узлы распределяются в нечеткое разбиение $P_{(\ell+1)}^i$ и в дальнейшем не подвергаются обработке GFI алгоритмом;
- 3 Если для всех l текущего уровня выполняется $P_{(\ell)}^i = P_{(\ell+1)}^i$, то алгоритм прекращает работу, в противном случае полагается $\ell := \ell + 1$ и осуществляется переход на шаг 2;
 - 4 Для каждого $A_{k(\ell)}^i$ атома разбиения $P_{(\ell)}^i$ при помощи GFI алгоритма вычисляется нечеткое разбиение $\{A_{1(\ell+1)}^i, A_{2(\ell+1)}^i\}$: если $A_{1(\ell+1)}^i$ и $A_{2(\ell+1)}^i$ являются «реальными» кластерами, то они образуют разбиение $\ell + 1$ -го уровня $P_{(\ell+1)}^i$, в противном случае атом $A_{k(\ell)}^i$ не образует разбиения $\ell + 1$ -го уровня $P_{(\ell+1)}^i$ и считается неделимым нечетким кластером — терминальным узлом, полагается $c := c + 1$ и осуществляется переход к рассмотрению другого атома.

Следует отметить, что, как отмечает Д. Думитреску [81, с.153], число терминальных узлов c в бинарном дереве соответствует «оптимальному» числу классов в исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, так что FDN алгоритм может использоваться для отыскания естественного числа классов перед применением оптимизационной кластер-процедуры, использующей число классов c в качестве входного параметра.

В работе [81] также предлагается адаптивная метрика для случая различного размера кластеров.

3.3.2.3. Другие алгоритмы иерархического подхода к решению нечеткой модификации задачи автоматической классификации, различные аспекты которого более подробно рассматриваются С. Мийамото [128], также немногочисленны. Среди этих алгоритмов особо следует отметить агломеративную процедуру, описываемую С. Ф. Боклишем [64], методы, основанные на выделении классов толерантности, предлагаемые И. З. Батыршиным [3], [4], [5], [6], [7], [8], и послужившие основой программной системы КЛАСТИЕР [10], а также иерархическую версию эвристического алгоритма Тамуры — Хигути — Танаки, основанную на результатах, полученных Дж. Данном [82], [86], так что иерархия разбиений на четкие классы представляется в виде дерева иерархии или дендрограммы [182, с.164].

Алгоритм Думитреску (FHDR procedure) [81, с.155-156] представляет собой модификацию FDH алгоритма для решения задачи снижения признакового пространства.

Основные понятия и определения

Если данные об исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$ представлены в виде матрицы «объект-свойство» $X_{m \times n} = [x'_i], i=1, \dots, n, t=1, \dots, m$, так что каждый объект рассматривается как точка в пространстве $I^m(X)$, то x'_i представляет собой значение t -го признака на i -м объекте. Таким образом, t -я строка $X' = (x'_1, x'_2, \dots, x'_n)$ матрицы вида (1.1) полностью характеризует признак x^t .

Среднее значение признака x^t для объектов класса A^t определяется выражением

$$v_{A^t}(x^t) = \frac{1}{card(A^t)} \sum_{i=1}^n \mu_i x'_i, \quad (3.127)$$

тогда вариация признака x^t в классе A^t может быть определена как

$$\sigma_{A^t}^2(x^t) = v_{A^t}((x^t - v_{A^t}(x^t))^2), \quad (3.128)$$

так что, очевидно, имеет место

$$\sigma_{A^t}^2(x^t) = v_{A^t}((x^t)^2) - (v_{A^t}(x^t))^2, \quad (3.129)$$

после чего значение признака x^t можно определить следующим образом:

$$\hat{x}^t = (x^t - v_{A^t}(x^t)) / \sigma_{A^t}(x^t). \quad (3.130)$$

Процедура снижения размерности признакового пространства $I^m(X)$ может рассматриваться как процесс классификации признаков $M = \{x^1, \dots, x^m\}$ после их преобразования в соответствии с формулой (3.130) с целью отбора $\hat{m} \leq m$ наиболее информативных признаков. Если классы признаков, характеризующие объекты множества $X = \{x_1, \dots, x_n\}$ являются однородными и хорошо разделенными, то каждый класс может быть представлен некоторым наиболее информативным признаком. Во избежание путаницы с процессом классификации объектов FDH алгоритмом при классификации признаков классы разбиения уровня ℓ будут обозначаться символом $B_{k(\ell)}^t$, а сами разбиения ℓ -го уровня — символом $F_{(\ell)}^t$.

Параметры алгоритма:

Как и в случае FDH алгоритма, исходные данные задаются в виде матрицы «объект-свойство» (1.1), а входные параметры отсутствуют. В процессе работы алгоритма используется счетчик $\hat{m}$.

Схема алгоритма:

- 1 Полагается $\ell := 0; \hat{m} := 0; F_{(0)} = \{M, \emptyset\}$;
- 2 Для текущего уровня ℓ нумеруются разбиения $l = 1, \dots$; для каждого разбиения $F'_{(l)}$ просматриваются все атомы, и для каждого непустого нечеткого кластера $B'_{k(l)} \in F'_{(l)}$ осуществляется выполнение условия, описанного на шаге 4; все неделимые кластеры $B'_{k(l)}$ распределяются в нечеткое разбиение $F'_{(\ell+1)}$ и в дальнейшем не подвергаются обработке GFI — алгоритмом;
- 3 Если для всех l текущего уровня выполняется $F'_{(l)} = F'_{(\ell+1)}$, то алгоритм прекращает работу, в противном случае полагается $\ell := \ell + 1$ и осуществляется переход на шаг 2;
- 4 Для каждого $B'_{k(l)}$ атома разбиения $F'_{(l)}$ при помощи GFI алгоритма вычисляется нечеткое разбиение $\{B'_{1(\ell+1)}, B'_{2(\ell+1)}\}$: если $B'_{1(\ell+1)}$ и $B'_{2(\ell+1)}$ являются «реальными» кластерами, то они образуют разбиение $\ell + 1$ -го уровня $F'_{(\ell+1)}$, в противном случае атом $B'_{k(l)}$ считается неделимым и полагается $\hat{m} := \hat{m} + 1$; полагается $\bar{x}^i = \bar{x}$, где все значения $\bar{x}^i$ вычисляются по формуле (3.130), а $\bar{x}$ — наиболее информативный признак класса $B'_{k(l)}$, так что $\mu_{k(l)}^i(x) = \max_{x' \in M} \mu_{k(l)}^i(x')$, и осуществляется переход к рассмотрению другого атома.

Таким образом, размерность исходного признакового пространства $I^m(X)$ оказывается сниженной до $\hat{m}$ и отобранные признаки образуют множество $\hat{M} = \{\bar{x}^1, \dots, \bar{x}^{\hat{m}}\}$.

Резюме

Нечеткие методы автоматической классификации эвристического направления представлены алгоритмом И. Гитмана и М. Д. Левина, исходные данные для которого представляются в форме матрицы «объект-свойство» алгоритмом, предложенным С. Тамурой, С. Хигути и К. Танакой, алгоритмом А. Кутюрье и Б. Фьолео и алгоритмом классификации на нечетких графах Л. С. Берштейна и Т. А. Дзюбы, использующими в качестве исходных данных матрицу «объект-объект». Эти процедуры непосредственно опираются на постановку задачи вы-

деления в исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, структура которой характеризуется нечеткостью, групп объектов $A^1, \dots, A^c$, являющихся нечеткими множествами с соответствующими функциями принадлежности $\mu_{il}, l = 1, \dots, c, i = 1, \dots, n$. Нечеткие оптимизационные методы автоматической классификации находят решение задачи $Q(P) \rightarrow \underset{P \in \Pi}{extr}$, где Π — множество всех возможных нечетких разбиений P исходного множества объектов X . В свою очередь, нечеткие множества $A^l, l = 1, \dots, c$, определенные на универсуме $X = \{x_1, \dots, x_n\}$, с соответствующими функциями принадлежности $\mu_{il}, l = 1, \dots, c, i = 1, \dots, n$, образуют нечеткое c -разбиение $P = \{A^1, \dots, A^c\}$ в смысле Э. Г. Распини, если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^c \mu_{il} = 1$. В зависимости от вида матрицы

исходных данных представленные оптимизационные методы условно объединяются в две группы: при задании исходных данных в виде матрицы «объект-объект» для решения задачи классификации применяются процедуры Э. Г. Распини, М. П. Уиндхема, М. Рубенса, Р. Н. Даве и С. Сена, находящие экстремум соответствующего функционала, а если исходные данные заданы в виде матрицы «объект-свойство», решение задачи классификации заключается в нахождении экстремума одного из функционалов, предложенных Дж. Беждеком и Дж. Данном, В. Педричем или В. Райтом. Общим для оптимизационных методов нечеткого подхода к решению задачи автоматической классификации ограничением является условие нечеткого разбиения Э. Г. Распини, а общим параметром является параметр c , определяющий число кластеров в искомом нечетком разбиении. Представлены также различные модификации некоторых из рассмотренных оптимизационных методов. Нечеткие иерархические кластер-процедуры представлены алгоритмом Д. Ватады, Х. Танаки и К. Асаи, в качестве исходных данных использующим матрицу сходства объектов (1.3) и допускающим как восходящую, так и нисходящую версии, и бинарным дивизимным алгоритмом Д. Думитреску, который предполагает представление исходных данных в форме матрицы «объект-свойство». Модификация алгоритма Д. Думитреску применима для решения задачи снижения признакового пространства.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕЧЕТКИХ МЕТОДОВ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

4.1. Содержательные аспекты сравнения нечетких методов автоматической классификации

4.1.1. Цели сравнения и оценка сходимости нечетких кластер-процедур

В существующих подходах к решению проблемы классификации объектов в условиях отсутствия обучающих выборок доля нечетких методов кластер-анализа составляет пока весьма незначительную часть от общего числа методов автоматической классификации. Как отмечал И. Д. Мандель, «мы не считаем язык теории размытых множеств особенно удобным для теории классификации. Он универсален так же, как и язык бинарных отношений и «обычных» мер близости объектов, а принципиально новые конструкции с его помощью не получаются. Но с точки зрения интерпретации полученных решений концепция размытых множеств весьма удобна...» [31, с.62-63]. То обстоятельство, что динамика развития нечетких методов автоматической классификации остается сравнительно невысокой, обусловлено следующими взаимосвязанными факторами.

Во-первых, теория нечетких множеств является новой областью математики, которая продолжает развиваться весьма бурными темпами. В силу этого обстоятельства использование как всего богатейшего арсенала, имеющегося в распоряжении теории нечетких множеств, так и ее последних достижений в задачах кластер-анализа продолжает оставаться незначительным.

Во-вторых, в силу вышеизложенного основные понятия нечеткой классификации являются недостаточно хорошо осмыслены и, как следствие, отсутствует единая общепринятая формулировка нечеткой модификации задачи автоматической классификации. Этим фактором обусловлено также и то, что существующие нечеткие методы кластерного анализа в подавляющем большинстве представляют собой, как показывают теоретические рассуждения, просто нечеткие аналогии, обобщения или, напротив, частные случаи традиционных алгорит-

мов автоматической классификации; в особенности это касается оптимизационных кластер-процедур [37, с.241], [31, с.77-96].

Таким образом, методология сравнительного анализа существующих алгоритмов нечеткой автоматической классификации не должна принципиально отличаться от методологии сравнения традиционных методов кластер-анализа. Однако перед рассмотрением схемы сравнения нечетких кластер-процедур целесообразно указать цели проводимого сравнения, поскольку любое сравнение кластер-процедур проводится в зависимости от цели этого сравнения. Собственно, цель сравнительного анализа обусловлена, как правило, либо теоретическими рассмотрениями, либо конкретным прикладным исследованием. В первом случае сравнительный анализ проводится с целью определения эффективности кластер-процедур для решения некоторого класса задач. Во втором случае, как правило, между собой сравниваются результаты работы нескольких алгоритмов для выявления реальной структуры исследуемой совокупности и последующего уточнения результатов.

Эффективность любой кластер-процедуры может оцениваться по уровню устойчивости к возмущениям, оценке трудоемкости, а также оценке сходимости алгоритма; более подробно эти вопросы рассматриваются И. Д. Манделем [31]. Касательно оптимизационных кластер-процедур нечеткого подхода следует отметить, что для оценки их сходимости может быть использовано соотношение [140]

$$e(b) = \max_{l,i} |\mu_{(b)li} - \mu_{(b-1)li}|, \quad (4.1)$$

где $\mu_{(b)li}$ — значение принадлежности i -го объекта l -му кластеру на b -м шаге вычислительного процесса, $b=1,2,\dots$, так что оказывается возможным построить график изменения величины $e(b)$, иллюстрирующий сходимость вычислительного процесса.

4.1.2. Виды исходных данных и общая схема проведения сравнительного анализа

В зависимости от характера проводимого исследования следует различать несколько видов исходных данных для сравнения работы нечетких кластер-процедур. Проблема экспериментального сравнения алгоритмов кластерного анализа достаточно подробно рассматривалась И. Д. Манделем [31, с.107-117], который выделял несколько подходов к ее решению.

Первый подход заключается в использовании реальных данных с неизвестной структурой и сравнении результатов работы различных алгоритмов между собой, а если в распоряжении исследователя имеется экспертное разбиение, то эти результаты сравниваются и с ним. Касательно данного подхода И. Д. Мандель отмечал, что «этот способ малоубедителен для общих выводов, но очень полезен в конкретном исследовании, где близость результатов почти всегда говорит о структурированности в данных» [31, с.107].

Второй подход состоит в использовании для тестирования алгоритмов реальных данных с известной структурой, на которых уже были опробованы различные алгоритмы классификации и которые могут послужить хорошим тестом для новых кластер-процедур. Примером подобных данных может послужить матрица четырехмерных данных по 150 ирисам Е. Андерсона [45]. Вместе с тем, число подобных наборов данных незначительно, кроме того, как подчеркивал И. Д. Мандель, «данные такого типа, безусловно, весьма привлекательны, но и они не могут убедительно ответить на вопрос о качестве алгоритмов в общем случае, так как успешное разбиение на конкретной выборке не гарантирует успеха на другой» [31, с.108].

Третий подход заключается в подборе искусственных данных с хорошо известной и определенной структурой, в которых кластеры могут быть выделены визуально, так что подобные массивы данных, как правило, являются двумерными. Этот подход предоставляет широкие возможности для сопоставления и осмысления человеческого и машинного способов классификации. Однако субъективизм человека при формировании данных является довольно сильным ограничением для применения такого подхода, так как «трудно установить устойчивые характеристики алгоритмов, связанные со случайным разбросом. К тому же, двумерная ситуация не является универсальной» [31, с.108].

Четвертый подход предполагает использование для экспериментального сравнения алгоритмов автоматической классификации генерируемых ЭВМ искусственных данных с заданной структурой, которые в процессе исследования будут искажаться контролируемым образом. Подобный способ является весьма предпочтительным и для проверки хорошо известных алгоритмов, поскольку позволяет моделировать данные самой разнообразной структуры и проводить управляемые искажения любого типа [31, с.108].

В рамках конкретного прикладного исследования производится классификация реальных данных, так что целью сравнения работы различных алгоритмов является, как правило, определение действительной структуры исследуемой совокупности для последующего детального рассмотрения. Таким образом, сравнительный анализ работы различных алгоритмов должен проводиться в зависимости от цели предпринимаемого исследования. Если целью сравнения является определение числа нечетких кластеров, то следует сравнить результаты работы одной или нескольких оптимизационных нечетких кластер-процедур при различном числе кластеров. В свою очередь, при обращении к оптимизационному подходу для отыскания истинной структуры данных следует использовать несколько различных функционалов. Поскольку в случаях использования разных функционалов классификации будут различаться, тогда если при использовании нескольких функционалов окажется, что, как указывал И. Д. Мандель, «...классификации похожи, то скорее всего выявлена реальная структура» [31, с.38]. Для выявления взаимосвязи между свойствами отдельных кластеров и разбиением в целом целесообразно сравнить между собой результаты работы оптимизационных и эвристических кластер-процедур, примером чего может послужить сравнение результатов работы алгоритма Беждека — Данна с результатами работы алгоритма Кутюрье — Фьолео при различных параметрах [71]. Если же необходимо определить характер взаимосвязи между стратификационной и кластерной структурами исследуемой совокупности объектов, то целесообразно сравнить результаты работы оптимизационных и иерархических кластер-процедур. К примеру, если исходные данные представлены в виде матрицы «объект-свойство», то для выявления подобной взаимосвязи следует сравнить результаты работы алгоритма Беждека — Данна и алгоритма Думитреску, а если исходные данные представлены в виде матрицы «объект- объект», то можно сравнить результаты работы, к примеру, алгоритма Уиндхема и алгоритма Ватады — Танаки — Асаи. Если же в процессе исследования возникает задача выявления взаимосвязи между стратификационной структурой исследуемой совокупности объектов и свойствами нечетких классов, то следует сравнить результаты работы нечетких иерархических и нечетких эвристических алгоритмов автоматической классификации. При необходимости детального исследования свойств нечетких кластеров целесообразно применить к исходным данным несколько эвристических нечетких кластер-процедур и сравнить между собой ре-

результаты их работы. Наконец, если целью исследования является получение наиболее полного представления о стратификационной структуре исследуемой совокупности, следует сравнить между собой результаты работы нескольких иерархических кластер-процедур.

Безусловно, предложенная схема, в силу предельной общности, не является окончательной и никоим образом не претендует на универсальность. Вместе с тем, следование предложенной установке позволит, более эффективно проводить сравнение алгоритмов как в процессе теоретического исследования, так и в процессе конкретного прикладного исследования.

Великолепным примером применения сравнения различных оптимизационных процедур для выявления характера и оценки кластерной структуры исследуемой совокупности объектов, может послужить методология, предложенная В. Педричем [144], сущность которой состоит в следующем.

При обработке исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$ несколькими оптимизационными процедурами, в каждом случае получается матрица разбиения $P_{c \times n}^b = [\mu_{ij}^b]$, $b = 1, \dots, K$, где K — число методов, которыми обрабатывалось множество объектов. Таким образом, оказывается возможным построить последовательность матриц $P_{c \times n}^1, P_{c \times n}^2, \dots, P_{c \times n}^K$, и i -й столбец любой из них выражает степени принадлежности i -го объекта каждому из c нечетких кластеров. Для выявления взаимосвязи между кластерами, полученными при использовании нескольких методов, необходимо сопоставить соответствующие строки матриц $P_{c \times n}^1, P_{c \times n}^2, \dots, P_{c \times n}^K$. Субоптимальная процедура, предложенная В. Педричем [144, с.137-142] для решения этой проблемы, может быть представлена в виде следующей последовательности шагов:

- 1 Полагается $b := 1, l := 1$;
- 2 Сравнивается l -я строка матрицы разбиения $P_{c \times n}^b$ со строками матрицы разбиения $P_{c \times n}^{b+1}$, и отыскивается индекс $l_0, l_0 = 1, \dots, c$, такой, что некоторое расстояние, к примеру, Хемминга, между l -й строкой матрицы $P_{c \times n}^b$ и l_0 -й строкой матрицы $P_{c \times n}^{b+1}$ достигает минимума;
- 3 Предыдущий шаг выполняется для всех $l, l = 1, \dots, c$, что позволяет построить функцию $i = i(l)$, выражающую соответствие между строками матриц разбиения $P_{c \times n}^b$ и $P_{c \times n}^{b+1}$;

4 Вычисляются средние значения функций принадлежности соответствующих кластеров, так что строится средняя матрица разбиения;

5 Если $b < K$, то полагается $b := b + 1$ и осуществляется переход на шаг 2, производится замещение матрицы разбиения $P_{\text{ср}}^b$ вычисленной средней матрицей разбиения; при $b = K$ процесс останавливается.

Результаты кластеризации несколькими методами могут быть представлены в виде таблицы для каждого кластера $A^l, l=1, \dots, c$, при-
мером чему может послужить таблица 4.1.

Таблица 4.1

Результат обработки данных несколькими нечеткими оптимизационными кластер-процедурами для одного кластера

Объект	Метод кластеризации			
	Метод 1	Метод 2	...	Метод K
x_1	μ_{11}^1	μ_{11}^2	...	μ_{11}^K
x_2	μ_{12}^1	μ_{12}^2	...	μ_{12}^K
...	...	...	...	...
x_n	μ_{1n}^1	μ_{1n}^2	...	μ_{1n}^K

Таким образом, речь идет о степенях принадлежности всех объектов исследуемой совокупности некоторому l -му кластеру, $l=1, \dots, c$. Каждый столбец подобной матрицы соответствует представлению l -го кластера, порожденного соответствующим методом кластеризации. Как отмечает далее В. Педрич [144, с.137-138], «мысленно «прокручивая» все таблицы для полученных кластеров, мы имеем дело с c вероятностными множествами, так что при таком подходе кластер оказывается более нечетким, а вероятностным множеством, тогда как при применении какого-либо одного метода он превращается в нечеткое множество».

Выражая сложность как некоторого объекта x_i относительно кластера A^l , так и всей исследуемой совокупности объектов $X = \{x_1, \dots, x_n\}$, в терминологии энтропии оказывается возможным оценить не только делимость нечетких кластеров, но и эффективность того или иного метода кластеризации.

4.2. Оценка и представление результатов обработки данных нечеткими методами автоматической классификации

4.2.1. Оценка качества нечеткой классификации

Как правило, результаты классификации, помимо содержательного осмысления и интерпретации, подлежат некоторой объективной оценке. В силу разнородности нечетких методов автоматической классификации, что особенно характерно для группы эвристических методов, каких-либо универсальных показателей качества полученной классификации не предлагается. В случае эвристических алгоритмов методы оценки классификации иногда предлагаются вместе с конкретным алгоритмом классификации. В частности, для оценки классификации С. Тамура, С. Хигути и К. Танака [163, с.64-65] используют такие показатели, как коэффициент корректно классифицированных объектов, коэффициент некорректно классифицированных объектов и коэффициент нерасклассифицированных объектов. Следует, однако, указать, что для вычисления подобных коэффициентов необходимо знать реальную структуру данных или, в крайнем случае, иметь в наличии экспертное разбиение, так что эти коэффициенты могут быть использованы для оценки собственно метода классификации, а не результатов его применения к некоторым данным с неизвестной структурой.

В отличие от эвристических методов нечеткого подхода к решению задачи кластерного анализа, для группы оптимизационных методов предложены несколько показателей, характеризующих полученное нечеткое разбиение $P^* = \{A^1, \dots, A^c\}$, которое описывается следующей матрицей:

$$P_{c \times n} = \begin{pmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \dots & \dots & \dots & \dots \\ \mu_{c1} & \mu_{c2} & \dots & \mu_{cn} \end{pmatrix}, \quad (4.2)$$

где, как и ранее, μ_{ii} — значение принадлежности элемента $x_i \in X$ некоторому нечеткому кластеру $A^i \in \{A^1, \dots, A^c\}$, n — количество элементов классифицируемого множества, или, иными словами, универ-

сума $X = \{x_1, \dots, x_n\}$, а c — число нечетких кластеров в полученном разбиении.

Наиболее известными из предложенных показателей является коэффициент разбиения, предложенный в работе [85] и подробно обсуждавшийся в работах [58], [166]:

$$F_c(P) = \frac{1}{n} \sum_{l=1}^c \sum_{i=1}^n \mu_{li}^2, \quad (4.3)$$

а также энтропия разбиения, предложенная в работе [58]:

$$H_c(P) = \frac{1}{n} \sum_{l=1}^c \sum_{i=1}^n |\mu_{li} \cdot \ln \mu_{li}|. \quad (4.4)$$

Данные показатели обладают следующими свойствами:

- 1) В случае, когда полученное разбиение является четким, то есть μ_{li} принимает значения на двухэлементном множестве $\{0,1\}$, характеризующем принадлежность i -го элемента l -му кластеру, $F_c(P) = 1$ и $H_c(P) = 0$;
- 2) В случае, когда полученное разбиение является наиболее неопределенным, то есть когда $\mu_{li} = \frac{1}{c}$ для всех $i=1, \dots, n$ и $l=1, \dots, c$, показатели принимают значения $F_c(P) = 1/c$ и, соответственно, $H_c(P) = \ln c$.

Таким образом, диапазон значений коэффициента разбиения определяется неравенством $1/c \leq F_c(P) \leq 1$, а диапазон значений энтропии разбиения — неравенством $0 \leq H_c(P) \leq \ln c$.

Главной целью использования коэффициента разбиения $F_c(P)$ и энтропии разбиения $H_c(P)$ «является отыскание наиболее «приемлемого» числа кластеров» [144, с.135] в нечетком разбиении P^* . Вместе с тем, необходимо отметить, что диапазон значений $F_c(P)$ и $H_c(P)$ зависит от числа нечетких кластеров c , то есть при изменении числа кластеров в нечетком разбиении соответствующим образом изменяются и значения обоих показателей, так что коэффициент разбиения $F_c(P)$ и энтропия разбиения $H_c(P)$ оказываются непригодными для сравнения разбиений с различным числом кластеров одной исследуемой совокупности. В. Педрич по этому поводу отмечал, что «минимальное или максимальное значение показателей может быть найдено, или, в крайнем случае, может наблюдаться существенный скачок их значений. Тем не менее их поведение не находит теоретического обоснования. Вычислительные эксперименты показывают, главным

образом, их полезность, однако невозможно судить, какое значение показателей является наилучшим и наиболее соответствующим числу классов» [144, с.135]. Таким образом, целесообразным оказывается использование *пропорциональной экспоненты*, предложенной в работе [184]:

$$E_c(P) = -\ln \prod_{i=1}^n \sum_{l=1}^{m_i} (-1)^{l+1} \binom{c}{l} \left(1 - l \cdot \sqrt[c]{\mu_{il}}\right)^{c-1}, \quad (4.5)$$

диапазон значений которой не зависит от числа нечетких кластеров в нечетком разбиении и выражается неравенством $0 \leq E_c(P) \leq \infty$.

Коэффициент разбиения $F_c(P)$, энтропия разбиения $H_c(P)$, а также другие показатели, характеризующие нечеткое разбиение, подробно исследуются в работах [58], [119]; кроме того, проблеме обоснования числа кластеров посвящены работы [52], [120]. Вместе с тем, некоторыми исследователями предлагаются либо модификации существующих показателей для оценки предлагаемых ими процедур, либо специальные методы оценки эффективности предлагаемой нечеткой кластер-процедуры. К примеру, М. Рубенсом [148, с.246] была предложена следующая модификация коэффициента разбиения $F_c(P)$:

$$NFI_c(P) = \frac{c \left[\sum_{l=1}^c \sum_{i=1}^n \mu_{il}^2 \right] - n}{n(c-1)}, \quad (4.6)$$

что можно переписать в виде $NFI_c(P) = \frac{cF_c(P) - 1}{c-1}$ [166, с.238].

Помимо модификации коэффициента разбиения $NFI_c(P)$, в этой же работе [148] М. Рубенсом была предложена также следующая оценка разбиения:

$$NFIP_c(P) = \frac{\sum_{l>l'} \delta(l, l')}{\binom{c}{2}}, \quad (4.7)$$

где $\delta(l, l')$ представляет индекс различия пары кластеров (l, l') , так

что $\delta(l, l') = 1 - \frac{\sum_{i=1}^n \min\{\mu_{il}, \mu_{il'}\}}{\sum_{i=1}^n \max\{\mu_{il}, \mu_{il'}\}}$. Как и для коэффициента разбиения

$F_c(P)$, для показателей разбиения М. Рубенса выполняются неравенства $1/c \leq NFI_c(P) \leq 1$ и $1/c \leq NFIP_c(P) \leq 1$.

Исследуя коэффициент разбиения $F_c(P)$, Е. Трауверт отмечает, что « $F_c(P)$ измеряет четкость, но не ее точные предельные значения, выраженные соотношением (4.3), а ряд ее полностью определенных значений» [166, с.221]. При рассмотрении взаимосвязи между наименьшим и наибольшим значениями $F_c(P)$ Е. Трауверт вводит коэффициент нечеткости $D_c(P)$ разбиения $P^* = \{A^1, \dots, A^c\}$, вычисляемый по формуле

$$D_c(P) = \|W - P^*\|^2 / n, \quad (4.8)$$

где W — матрица четкого разбиения, имеющая, так же, как и матрица P^* , размерность $c \times n$:

$$W_{c \times n} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{c1} & w_{c2} & \dots & w_{cn} \end{pmatrix}, \quad (4.9)$$

элементы которой определяются следующим образом:

$$\left. \begin{aligned} w_{i_0 i} &= 1 \Leftrightarrow \mu_{i_0 i} = \max_{l=1}^c \mu_{li} \\ w_{li} &= 0 \Leftrightarrow l \neq i_0 \end{aligned} \right\} \forall i, \quad (4.10)$$

$$\|W - P^*\| = \sqrt{\sum_{l=1}^c \sum_{i=1}^n (w_{li} - \mu_{li})^2}. \quad (4.11)$$

Взаимосвязь между коэффициентом четкости $F_c(P)$ и коэффициентом нечеткости $D_c(P)$, в свою очередь, определяется выражением

$$F_c(P) = D_c(P) - 1 + \frac{2}{n} \sum_{l=1}^c \sum_{i=1}^n w_{li} \mu_{li}. \quad (4.12)$$

Используя методологию, детально рассмотренную в работе [166], можно построить соответствующий график, называемый Е. Траувертом (D, F) -диаграммой, схематический пример которой приведен на рис. 4.1.

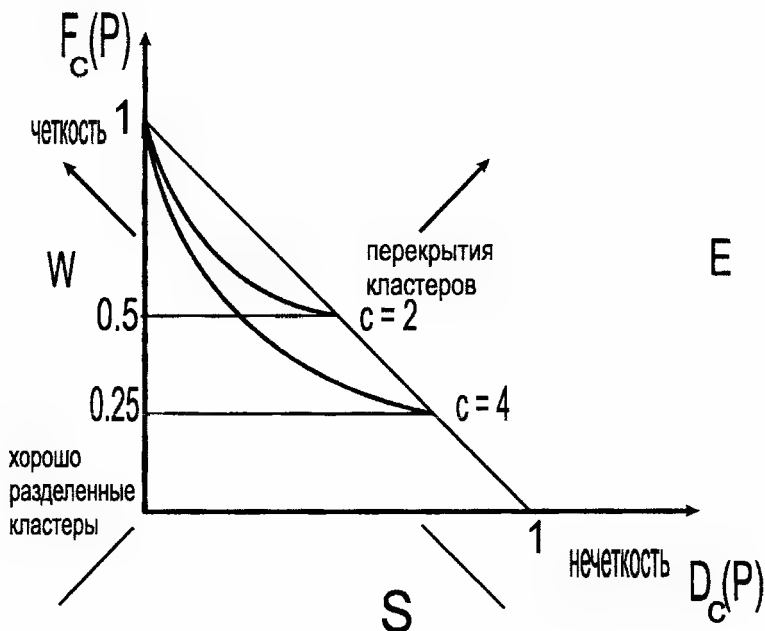


Рис. 4.1. Диаграмма Траувейта и ее интерпретация

Представление матрицы нечеткого разбиения P^* с помощью диаграммы Траувейта соответствует среднему всех представлений векторов принадлежности, составляющих матрицу P^* .

Интерпретация диаграммы Траувейта может проводиться в двух направлениях, как это изображено на рис. 4.1:

- 1) по направлению от $S-E$ к $N-W$ происходит уменьшение нечеткости и возрастает четкость разбиения;
- 2) по направлению от $S-W$ к $N-E$ происходит следующее изменение: от матриц разбиения, выражающих принадлежность каждого объекта в основном только одному кластеру, с некоторыми несущественными значениями принадлежности ко всем остальным кластерам, к матрицам, выражающим, что некоторые объекты в равной степени связаны несколькими кластерами.

Последняя ситуация известна также, как случай «цепочек», «мостиков» между кластерами и схематически изображена на рис. 1.4 в виде кластеров А и В.

Е. Трауверт отмечал, что « (D, F) -диаграмма проявляет себя как интересное новое средство, характеризующее задачу кластер-анализа или алгоритм нечеткой кластеризации, но не выбирающее обоснование нечеткого решения» [166, с.235].

Предпринятое рассмотрение различных методов оценки нечеткой классификации наглядно демонстрирует, что показатели, характеризующие нечеткое разбиение, несмотря на их полезность для анализа полученной классификации, не могут быть использованы для выявления действительной структуры исследуемой совокупности объектов, так что нельзя не согласиться с В. Педричем, отмечавшим, что при решении задачи нечеткой классификации «будет разумным использовать различные методы и сравнивать полученные результаты» [144, с.135]. Соображения, подобные приведенному замечанию, просматриваются также в работах [96], [97], [98].

В свою очередь, помимо оценки качества разбиения, можно также исследовать некоторые характеристики отдельных кластеров для детального рассмотрения их структуры. К примеру, Э. Г. Распини [149], [150] для исследования свойств отдельных кластеров использовал уже упоминавшееся соотношение

$$S_l = \sum_{j=1}^n p(x_j) \mu_{lj}, \quad (4.13)$$

выражающее относительный размер кластера $A^l, l \in \{1, \dots, c\}$, а также соотношение

$$P(x_i) = \sum_{j=1}^n p(x_j) d(x_i, x_j), \quad (4.14)$$

показывающее среднюю плотность точек вокруг некоторой точки $x_i \in X$. Кроме того, для оценки средней плотности точек вокруг точки x_i в нечетком кластере $A^l, l \in \{1, \dots, c\}$ Э. Г. Распини предложил также соотношение

$$P_l(x_i) = \frac{\sum_{j=1}^n p(x_j) \mu_{lj} d(x_i, x_j)}{\sum_{j=1}^n p(x_j) \mu_{lj}}. \quad (4.15)$$

В. Педричем [144] предлагается весьма интересный и многообещающий подход к решению проблемы представления структуры

нечетких кластеров, так что целесообразно его рассмотреть подробнее. Поскольку значение функции μ_{li} , представляющее собой элемент матрицы разбиения $P_{c \times n} = [\mu_{li}], l=1, \dots, c, i=1, \dots, n$, выражает степень принадлежности i -го элемента l -му кластеру, то элемент $x_i \in X$ может рассматриваться как элемент ядра нечеткого кластера A^l , если, по меньшей мере, это значение μ_{li} превышает некоторый порог, который в дальнейшем будет обозначаться символом ϕ . Далее, рассматривая для каждой точки $x_i \in X$ ее структурное отношение ко всем кластерам $A^l, l=1, \dots, c$, принимают во внимание функции принадлежности точки x_i всем остальным нечетким кластерам, за исключением рассматриваемого нечеткого кластера A^l . Поскольку число кластеров равно c , можно отметить, что любая точка $x_i \in X$, для которой $\mu_{li} = \frac{1}{c}, l=1, \dots, c$, не способствует выявлению сущности кластерной структуры. Мерой структурного свойства некоторой точки $x_i \in X$ может послужить показатель

$$\xi'(x_i) = -\prod_{l=1}^c \mu_{li} + \frac{1}{c^c}, \quad (4.16)$$

так что, производя нормализацию этого показателя умножением правой части выражения (4.16) на константу c^c , можно получить

$$\xi(x_i) = 1 - c^c \prod_{l=1}^c \mu_{li}, \quad (4.17)$$

то есть $\xi: X \rightarrow [0,1]$, и из выражения (4.17) и свойств матрицы разбиения $P_{c \times n} = [\mu_{li}], l=1, \dots, c, i=1, \dots, n$ следует, что если точка x_i принадлежит одному из кластеров $A^l, l=1, \dots, c$ со степенью принадлежности, равной 1, то значение показателя (4.17) в этой точке достигает максимума, то есть $\xi(x_i) = 1$. Структурная сущность ядра нечеткого кластера может определяться некоторым пороговым значением ϕ , так что точка x_i может быть элементом ядра нечеткого кластера, только если $\xi(x_i)$ превышает ϕ . Комбинация этих рассуждений приводит к определению понятия (ϕ, φ) -ядра нечеткого кластера $A^l, l=1, \dots, c$ как подмножества X_{φ}^l универсума $X = \{x_1, \dots, x_n\}$ при $\phi, \varphi \in [0,1]$, если оно содержит множество $\left\{ x_i \in X \mid 1 - c^c \prod_{l=1}^c \mu_{li} \geq \phi, \mu_{li} \geq \varphi \right\}$, то есть

$$X_{\varphi\varphi}^l = \{x_i \in X \mid \xi(x_i) \geq \phi, \mu_{li} \geq \varphi\}. \quad (4.18)$$

Таким образом, получается множество ядер $X_{\varphi\varphi}^l, l=1, \dots, c$ и резидуальное множество данных X^R , содержащее все оставшиеся элементы универсума X , так что выполняется соотношение

$$X = \bigcup_{l=1}^c X_{\varphi\varphi}^l \cup X^R, \quad (4.19)$$

где первая составляющая правой части представляет собой существенную для рассмотрения структуру, а вторая соответствует малосущественной структуре множества X .

Для аналитического представления (ϕ, φ) -ядра нечеткого кластера $A^l, l=1, \dots, c$ в N -мерном пространстве целесообразно рассматривать гиперболы, описываемые, в общем, выражением

$$(x_i - \tau^l)^T (x_i - \tau^l) = R_l^2, i=1, \dots, n, \quad (4.20)$$

где символ T обозначает операцию транспонирования, так что центрами гипербол будут точки τ^l , а радиусами, соответственно, значения R_l , для всех $l=1, \dots, c$. При рассмотрении l -й гиперболы точка τ^l , как точка, являющаяся прототипом l -го кластера, может быть опущена, либо она может трактоваться как некоторая точка $x_i \in X$, такая, что функция принадлежности μ_{li} l -му кластеру для нее является максимальной. Значение R_l выбирается таким образом, что положение точек, являющихся элементами ядра $X_{\varphi\varphi}^l$, минимизирует взвешенную сумму квадратов ошибок, которая определяется выражением

$$Q(R_l) = \sum_{x_i \in X_{\varphi\varphi}^l} \left\{ \sum_{b=1}^N \frac{(x_{ib} - \tau_b^l)^2}{R_{lb}^2} - 1 \right\}^2 \mu_{li}. \quad (4.21)$$

При взятии первой производной $Q(R_l)$ для соответствующего R_l решение системы уравнений

$$\frac{\partial Q(R_l)}{\partial R_l} = 0, l=1, \dots, c \quad (4.22)$$

дает решение

$$R_l^2 = \sum_{x_i \in X_{\varphi\varphi}^l} \mu_{li} Z_i^2 / \sum_{x_i \in X_{\varphi\varphi}^l} \mu_{li} Z_i, \quad (4.23)$$

где

$$Z_i = \sum_{b=1}^N (x_{ib} - \tau_b^l)^2. \quad (4.24)$$

Налагая дополнительные ограничения в форме системы уравнений

$$\forall_{\substack{1 \leq l, f \leq c \\ l \neq f}} d(\tau^l, \tau^f) > R_l + R_f, \quad (4.25)$$

где $d(\tau^l, \tau^f)$ устанавливает евклидово расстояние между точками τ^l и τ^f , так что, варьируя значения ϕ и/или φ , оказывается возможным определить невыделенные гиперсферы, представляющие ядра. Более того, следует отметить, что при фиксированном параметре φ для каждой пары $\phi_1, \phi_2 \in [0, 1]$, упорядоченных так, что $\phi_1 > \phi_2$, соответствующие радиусы будут упорядочены как $R_l(\phi_1) \leq R_l(\phi_2)$. С точки зрения статистического подхода, гиперсферическая форма кластеров соответствует ситуации, когда кластеры описываются средним значением τ^l и ковариационной матрицей

$$C_l = \begin{pmatrix} R_1^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & R_2^2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & R_l^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & R_c^2 \end{pmatrix}. \quad (4.26)$$

Таким образом, построение аналитического представления ядер $X_{\varphi\varphi}^l, l=1, \dots, c$ позволяет идентифицировать резидуальное множество данных X^R .

Наряду с гиперсферической формой, в работе [144] также рассматривается ситуация гиперэллиптической формы ядер нечетких кластеров, рассуждения для которой незначительно отличаются от приведенных выше. Порог φ для каждого кластера $A^l, l=1, \dots, c$ обычно соответствует среднему значению функции принадлежности точек кластеру, так что

$$n\varphi_l = \sum_{i=1}^n \mu_{li}, l=1, \dots, c. \quad (4.27)$$

Следует отметить, что величина, определяемая выражением (4.27), соответствует мощности нечеткого множества $A^l, l=1, \dots, c$ [34, с.67]. При таком выборе порога центром ядра некоторого нечеткого кластера будет точка, имеющая максимальное значение принадлежности этому кластеру, так что во внимание будут приниматься только те точки, значения принадлежности которых превышают средние, так

что в результате ядра нечетких кластеров автоматически устанавливаются распределением значений принадлежностей. Вместе с тем, как указывает В. Педрич, порог φ может выбираться произвольно.

4.2.2. Представление и интерпретация результатов нечеткой классификации

После проведения классификации следует представить результаты проведенных экспериментов в форме, позволяющей подвергнуть эти результаты дальнейшему анализу. Данный вопрос достаточно подробно рассматривается И. Д. Манделем [31, с.159-161] и С. А. Айвазяном [37, с.311-330], так что ниже приводятся только некоторые приемы, которые могут оказаться полезными при интерпретации результатов нечеткой классификации.

В первую очередь, при использовании оптимизационных и эвристических нечетких кластер-процедур, следует рассматривать матрицу нечеткого распределения объектов по классам. Слово «распределение» подчеркивает, что результатом классификации не всегда будет нечеткое разбиение, так что матрица нечеткого разбиения будет являться результатом работы только оптимизационных процедур, а в случае использования, к примеру, алгоритма Кутюрье — Фьолео, результатом проведенного исследования будет нечеткое покрытие. Поскольку данный алгоритм допускает пересечение нечетких классов, то при его использовании в качестве результата классификации целесообразно предоставлять в распоряжение исследователя также матрицу пересечений нечетких классов.

Для оптимизационных процедур в качестве результатов должны также выводиться значения показателей, характеризующих полученное нечеткое разбиение, к примеру, значения коэффициента разбиения, энтропии разбиения, пропорциональной экспоненты и других показателей. Весьма полезным приемом будет построение графика поведения этих показателей, к примеру, диаграммы Трауверта. В ряде случаев, к примеру, при обработке данных несколькими оптимизационными кластер-процедурами в процессе сравнительного анализа, исследователю необходимо рассматривать сходимость алгоритмов, так

что оказывается целесообразным построение соответствующего графика.

В распоряжение исследователя необходимо предоставлять также визуальное представление результатов полученной классификации, что позволит быстро оценить размытость кластеров и выявить их центры. В качестве подобного средства, помимо традиционных, широко известных методов наглядного представления результатов [37, с.332-382], можно также предложить так называемую линейную диаграмму, на которой по оси ординат откладываются значения принадлежности, а по оси абсцисс — номера объектов, так что принадлежность объекта классу обозначается точкой, находящейся на пересечении линии, соответствующей номеру объекта, с линией, которая соответствует степени принадлежности объекта классу, номер которого указывается рядом с точкой.

В качестве иллюстративного примера рассмотрим матрицу нечеткого разбиения шести объектов на два нечетких кластера, представленную таблицей 4.2.

Таблица 4.2

Матрица нечеткого разбиения совокупности из шести объектов на два класса

Номер кластера	Номер объекта					
	1	2	3	4	5	6
1	0.1	0.2	1.0	0.6	0.8	0.6
2	0.9	0.8	0.0	0.4	0.2	0.4

Линейная диаграмма представленного нечеткого разбиения изображена на рис. 4.2. Следует также отметить, что линейная диаграмма может применяться для наглядного представления как нечеткого разбиения, так и нечеткого покрытия; более того, возможно построение трехмерной линейной диаграммы, где по оси ординат откладываются номера кластеров, а значения принадлежности — по оси аппликат. Вместе с тем, необходимо подчеркнуть, что линейная диаграмма может служить средством только предварительного анализа полученных результатов, а также отметить то обстоятельство, что она применима для анализа результатов обработки исследуемой совокупности сравнительно небольшого объема.

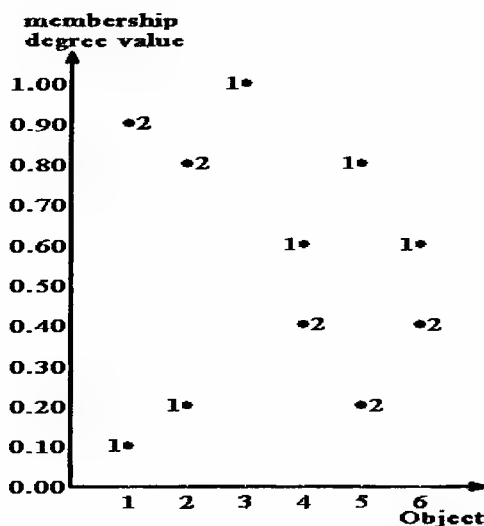


Рис. 4.2. Линейная диаграмма нечеткого разбиения исследуемой совокупности из шести объектов на два класса

При выборе в качестве инструмента исследования алгоритма классификации на нечетких графах Берштейна — Дзюбы результатом работы процедуры будет представление нечетких двудольных частей графа, а при использовании иерархических методов нечеткого подхода к решению задачи автоматической классификации в качестве результатов классификации исследователю должна предоставляться, естественно, дендрограмма или дерево нечеткой иерархии.

4.3. Экспериментальное сравнение нечетких методов автоматической классификации

4.3.1. Исходные данные для проведения экспериментального сравнения алгоритмов

С гносеологической точки зрения, качество машинной классификации определяется ее соответствием классификации, проведенной человеком, так что при значительной «близости» человеческой и машинной классификации можно сделать самые общие выводы о корректности той или иной процедуры. Данное обстоятельство диктует

необходимость использования при сравнении алгоритмов искусственных данных с известной структурой. С другой стороны, оценить эффективность кластер-процедур при помощи только таких данных довольно затруднительно. Охарактеризовать алгоритмы классификации можно при их тестировании на реальных данных, так что имеет смысл обратиться к реальным данным с известной структурой.

Целью предпринятого в данном случае сравнительного анализа является сопоставление результатов работы некоторых нечетких методов автоматической классификации с классификацией, проводимой человеком, для чего вполне достаточным является проведение сравнения результатов работы кластер-процедур с экспертным разбиением искусственных данных с известной структурой.

Подобный подход к проведению сравнения обуславливает некоторые требования к процессу формирования данных, используемых для сравнения. Поскольку сущность сравнительного анализа в данном случае заключается в определении эффективности алгоритма на основании сопоставления результатов его работы с человеческой классификацией, то, с учетом того обстоятельства, что человек не может классифицировать большое количество объектов, описанных также большим количеством признаков, первое требование выдвигается к числу объектов и размерности признакового пространства. В силу приведенных соображений число классифицируемых объектов и признаков должно быть небольшим, чтобы эксперт мог составить разбиение, соответствующее реальной структуре исследуемой совокупности, однако, учитывая замечание И. Д. Манделя об отсутствии универсальности двумерной структуры, число признаков должно быть большим чем три. Небольшое количество объектов обладает также тем методическим преимуществом, что позволяет сравнивать алгоритмы, работа которых затрудняется при больших объемах данных, к примеру, иерархических кластер-процедурах и процедурах классификации на графах. Второе требование касается «реалистичности» используемых данных с целью избежать человеческого субъективизма при их формировании, что, в свою очередь, диктует необходимость подбора в качестве исходных данных характеристик реальных объектов.

Учитывая приведенные требования, в качестве данных для экспериментального сравнения были отобраны одиннадцать самолетов различного предназначения [43], а в качестве описывающих их признаков были отобраны пять достаточно информативных характеристик, позволяющих построить экспертное разбиение человеку, не яв-

ляющемуся специалистом в области авиации. Эти данные представлены в таблице 4.3.

Таблица 4.3

Некоторые характеристики самолетов ВВС США

№	Тип самолета	Летно-технические данные				
		Размах крыла, м	Длина, м	Высота, м	Номинальная взлетная масса, кг	Максимальный радиус действия, км
1	F-104G	6.68	16.69	4.11	9428	1200
2	F-5A	7.70	14.68	4.06	6080	898
3	F-5E	8.13	14.68	4.06	7030	1083
4	F-16A	9.45	14.52	5.01	10205	925
5	F-15A	13.05	19.43	5.63	18824	1050
6	F-4B	11.70	17.76	4.96	20865	1600
7	F-4E	11.77	19.19	5.02	18818	1266
8	F-106A	11.67	21.56	6.18	16100	920
9	RA-5C	16.15	23.11	5.92	30300	1600
10	SR-71A	16.95	32.74	5.64	63505	1930
11	B-58A	17.32	29.49	9.53	68000	2600

Данные самолеты американской разработки имеют следующее назначение: модификация самолета F-104 «Старфайтер» F-104G и самолет F-16A «Файтинг Фолкон» являются многоцелевыми истребителями; модификация F-4E самолета F-4 «Фантом II» представляет собой истребитель дальнего проникновения; модификация F-5A самолета F-5 «Тайгер II», самолеты F-15A «Игл» и F-106A «Дельта Дарт» представляют собой истребители-перехватчики, а такие самолеты, как F-5E и F-4B, являющиеся модификациями F-5 и F-4 соответственно, представляют собой истребители-бомбардировщики; RA-5C «Виджилент» является палубным самолетом-разведчиком, SR-71A является стратегическим самолетом-разведчиком, а самолет B-58A «Хастлер» является стратегическим бомбардировщиком. Вместе с тем, представленные в таблице 4.3 технические характеристики позволяют условно сгруппировать эти самолеты в три класса: класс 1 — легкие самолеты — F-104G, F-5A, F-5E, F-16A; класс 2 — средние самолеты — F-15A, F-4B, F-4E, F-106A, RA-5C; класс 3 — тяжелые самолеты — SR-71A и B-58A. Таким образом, обозначая принадлежность объекта классу единицей, а ее отсутствие — нулем, оказывается возможным построить экспертное разбиение. Матрица экспертного разбиения представлена таблицей 4.4.

Экспертное разбиение исследуемой совокупности объектов на три класса

Номер класса	Номер объекта										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	0	0	0	0	0	0	0
2	0	0	0	0	1	1	1	1	1	0	0
3	0	0	0	0	0	0	0	0	0	1	1

При обозначении размаха крыла символом $\hat{x}^1$, длины — символом $\hat{x}^2$, высоты — символом $\hat{x}^3$, номинальной взлетной массы — символом $\hat{x}^4$, максимального радиуса действия — символом $\hat{x}^5$, а самих самолетов — символами $x_i, i=1, \dots, 11$, где символ i соответствует номеру самолета в таблице 4.3, и нормировке получившихся данных по формуле $x'_i = \hat{x}_i / \max \hat{x}_i$ получится матрица $X_{\text{норм}} = [x'_i], i=1, \dots, 11, t=1, \dots, 5$ «объект-признак» вида (1.1). При ее транспонировании она может быть представлена таблицей 4.5.

Таблица 4.5

Матрица нормированных данных

Объект	Признак				
	x^1	x^2	x^3	x^4	x^5
x_1	0.385	0.509	0.431	0.138	0.461
x_2	0.444	0.448	0.426	0.089	0.345
x_3	0.469	0.448	0.426	0.103	0.416
x_4	0.545	0.443	0.525	0.150	0.355
x_5	0.753	0.593	0.590	0.276	0.403
x_6	0.675	0.542	0.520	0.306	0.615
x_7	0.679	0.586	0.526	0.276	0.486
x_8	0.673	0.658	0.648	0.236	0.353
x_9	0.932	0.705	0.621	0.445	0.615
x_{10}	0.978	1.000	0.591	0.933	0.742
x_{11}	1.000	0.900	1.000	1.000	1.000

При таком подходе каждый объект $x_i, i=1, \dots, 11$, представляя собой точку в пятимерном пространстве признаков, может быть интерпретирован как нечеткое множество на универсуме $x', t=1, \dots, 5$

признаков, так что каждое значение $x'_i, i=1, \dots, 11, t=1, \dots, 5$ может быть представлено в виде функции принадлежности $\mu_{x'_i}(x')$, которая показывает степень выраженности t -го признака у i -го объекта. Таким образом, оказывается возможным построить матрицу попарных расстояний $d_{\text{поп}}$ между объектами, применяя к матрице «объект-признак», представленной таблицей 4.5, относительное обобщенное расстояние Хемминга (2.13) или относительное евклидово расстояние (2.14).

Для иллюстрации работы алгоритмов в рассматриваемом случае достаточно применения какого-либо одного вида расстояния, для чего было использовано относительное евклидово расстояние, которое в данном случае может быть определено выражением

$$d_E(x_i, x_j) = \sqrt{\frac{1}{m} \sum_{t=1}^m (\mu_{x'_i}(x'_t) - \mu_{x'_j}(x'_t))^2}, i, j=1, \dots, 11, m=5, \quad (4.28)$$

так что матрица попарных расстояний, оказывающаяся в данной ситуации матрицей нечеткого отношения несходства I , будет представлена таблицей 4.6.

Таблица 4.6

Матрица попарных расстояний между объектами

I	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	0.000										
x_2	0.068	0.000									
x_3	0.053	0.034	0.000								
x_4	0.100	0.069	0.065	0.000							
x_5	0.195	0.190	0.178	0.132	0.000						
x_6	0.170	0.195	0.168	0.154	0.108	0.000					
x_7	0.155	0.167	0.147	0.119	0.057	0.062	0.000				
x_8	0.186	0.183	0.177	0.130	0.060	0.143	0.088	0.000			
x_9	0.313	0.329	0.308	0.276	0.154	0.156	0.162	0.190	0.000		
x_{10}	0.515	0.545	0.526	0.503	0.390	0.378	0.390	0.412	0.262	0.000	
x_{11}	0.614	0.648	0.626	0.596	0.490	0.467	0.492	0.507	0.358	0.222	0.000

После применения к матрице попарных расстояний операции дополнения нечеткого отношения (2.31) получается нечеткая толерантность T , матрица которой представлена таблицей 4.7.

Матрица сходства между объектами

T	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1.000										
x_2	0.932	1.000									
x_3	0.947	0.966	1.000								
x_4	0.900	0.931	0.935	1.000							
x_5	0.805	0.810	0.822	0.868	1.000						
x_6	0.830	0.805	0.832	0.846	0.892	1.000					
x_7	0.845	0.833	0.853	0.881	0.943	0.938	1.000				
x_8	0.814	0.817	0.823	0.870	0.940	0.857	0.912	1.000			
x_9	0.687	0.671	0.692	0.724	0.846	0.844	0.838	0.810	1.000		
x_{10}	0.485	0.455	0.474	0.497	0.610	0.622	0.610	0.588	0.738	1.000	
x_{11}	0.386	0.352	0.374	0.404	0.510	0.533	0.508	0.493	0.642	0.778	1.000

Таким образом, таблицы 4.5 — 4.7 представляют все основные типы матриц исходных данных, которые будут использоваться в процессе сравнительного анализа.

4.3.2. Результаты вычислительных экспериментов

Для проведения сравнения были отобраны оптимизационные процедуры Беждека — Данна, Педрича и Уиндхема, иерархический вариант алгоритма Тамуры — Хигути — Танаки и алгоритм классификации на нечетких графах Берштейна — Дзюбы. Для всех оптимизационных процедур эксперименты проводились при числе классов, равном двум, трем и четырем, для процедуры классификации на нечетких графах Берштейна — Дзюбы эксперименты проводились при числе классов, равном двум и трем. При проведении вычислительных экспериментов с оптимизационными процедурами в качестве матрицы первоначального разбиения в каждом эксперименте для определенного числа классов использовалась одна и та же, сгенерированная случайным образом с соблюдением условия нечеткого разбиения Распини $\sum_{l=1}^c \mu_{li} = 1, \forall x_i \in X, l = 1, \dots, c$ матрица $P_{con} = [\mu_{li}]$. При обработке данных алгоритмом Беждека — Данна значение показателя нечеткости γ полагалось равным двум.

В таблице 4.8 приведены результаты разбиения исследуемой совокупности объектов алгоритмом Бежdeka — Данна на два класса. Линейная диаграмма разбиения приведена на рис. 4.3.

Таблица 4.8

Значения принадлежности объектов классам при разбиении исследуемой совокупности алгоритмом Бежdeka — Данна на два класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.039	0.037	0.028	0.016	0.048	0.061	0.021	0.050	0.407	0.950	0.949
2	0.961	0.963	0.972	0.974	0.952	0.939	0.979	0.950	0.593	0.050	0.051

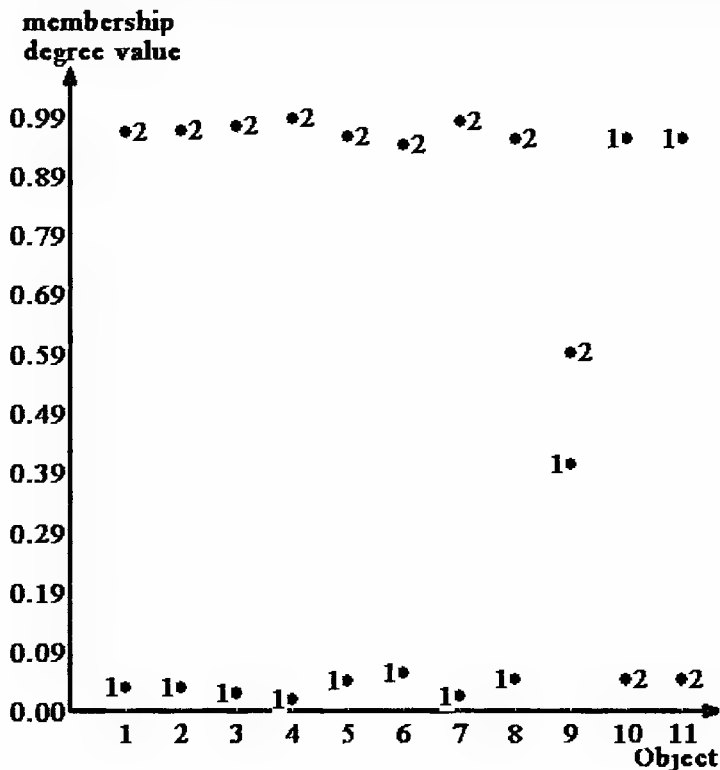


Рис. 4.3. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Бежdeka — Данна на два класса

Результаты разбиения исследуемой совокупности объектов алгоритмом Беждека — Данна на три класса приведены в таблице 4.9, а линейная диаграмма представлена на рис. 4.4.

Таблица 4.9

Значения принадлежностей объектов классам при разбиении исследуемой совокупности алгоритмом Беждека — Данна на три класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.925	0.970	0.988	0.868	0.054	0.166	0.061	0.190	0.140	0.043	0.028
2	0.008	0.003	0.001	0.009	0.008	0.027	0.006	0.025	0.149	0.871	0.924
3	0.067	0.027	0.011	0.123	0.938	0.807	0.933	0.785	0.711	0.086	0.048

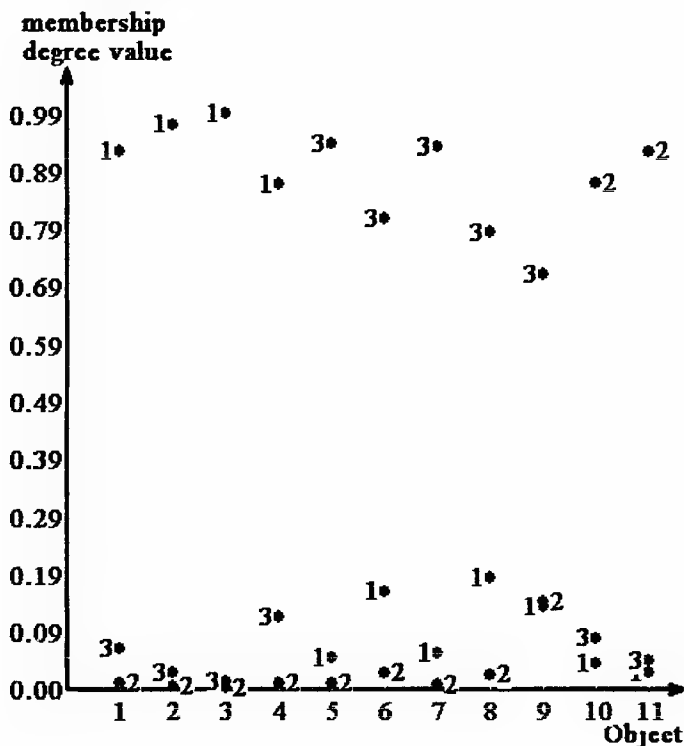


Рис. 4.4. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Беждека — Данна на три класса

Матрица разбиения исследуемой совокупности объектов алгоритмом Бежdeka — Дaнна на четыре класса представлена таблицей 4.10, а соответствующая линейная диаграмма — на рис. 4.5.

Таблица 4.10

Значения принадлежностей объектов классам при разбиении исследуемой совокупности алгоритмом Бежdeka — Дaнна на четыре класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.006	0.002	0.001	0.008	0.005	0.024	0.004	0.015	0.001	0.617	0.937
2	0.022	0.008	0.003	0.033	0.040	0.176	0.026	0.087	0.994	0.231	0.035
3	0.896	0.961	0.985	0.781	0.035	0.157	0.035	0.117	0.001	0.055	0.011
4	0.076	0.029	0.011	0.178	0.920	0.643	0.935	0.781	0.004	0.097	0.017

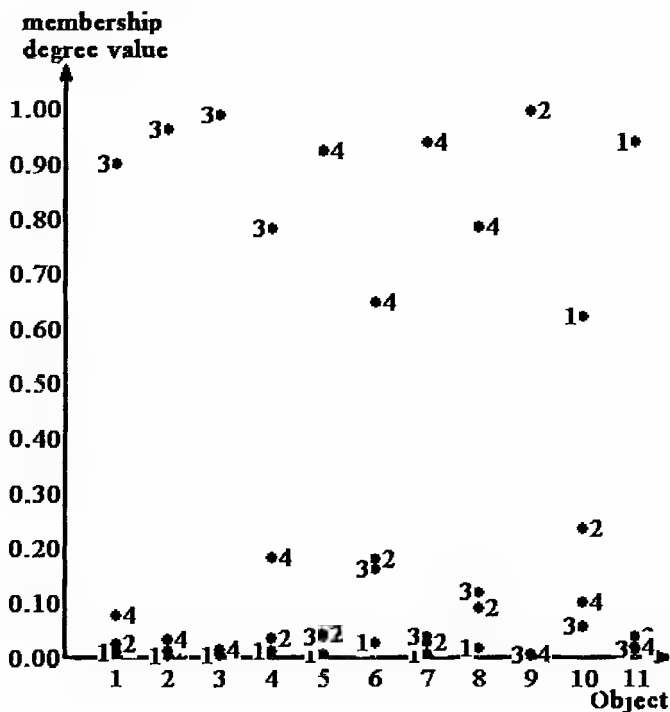


Рис. 4.5. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Бежdeka — Дaнна на четыре класса

В таблице 4.11 приведены результаты обработки исследуемой совокупности объектов алгоритмом Педрича, где в качестве помеченных объектов рассматривались объекты x_3 и x_{10} с функциями принадлежности $\mu_{13} = 1$ и $\mu_{210} = 1$ соответственно. Линейная диаграмма разбиения приведена на рис. 4.6

Таблица 4.11

Значения принадлежностей объектов классам при разбиении исследуемой совокупности алгоритмом Педрича на два класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.702	0.928	0.940	0.895	0.699	0.167	0.627	0.797	0.201	0.207	0.352
2	0.298	0.072	0.060	0.105	0.301	0.833	0.373	0.203	0.799	0.793	0.648

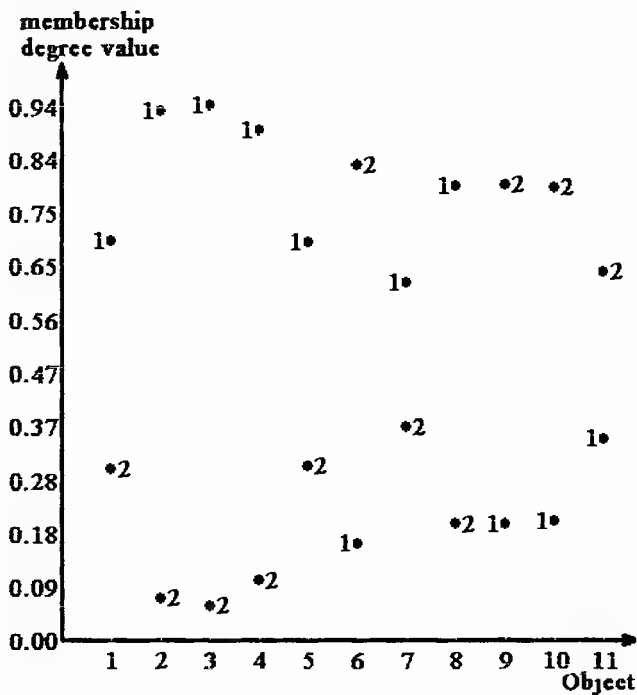


Рис. 4.6. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Педрича на два класса

Матрица разбиения исследуемой совокупности объектов алгоритмом Педрича на три класса, где в качестве помеченных объектов рассматривались объекты x_3 , x_6 и x_{11} с функциями принадлежности $\mu_{13}=1$, $\mu_{26}=1$ и $\mu_{311}=1$, представлена таблицей 4.12, а соответствующая линейная диаграмма — на рис. 4.7.

Таблица 4.12

Значения принадлежности объектов классам при разбиении исследуемой совокупности алгоритмом Педрича на три класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.644	0.943	0.942	0.598	0.230	0.109	0.021	0.478	0.093	0.066	0.025
2	0.253	0.048	0.049	0.347	0.697	0.826	0.973	0.420	0.810	0.129	0.047
3	0.103	0.009	0.009	0.055	0.073	0.065	0.006	0.102	0.097	0.805	0.928

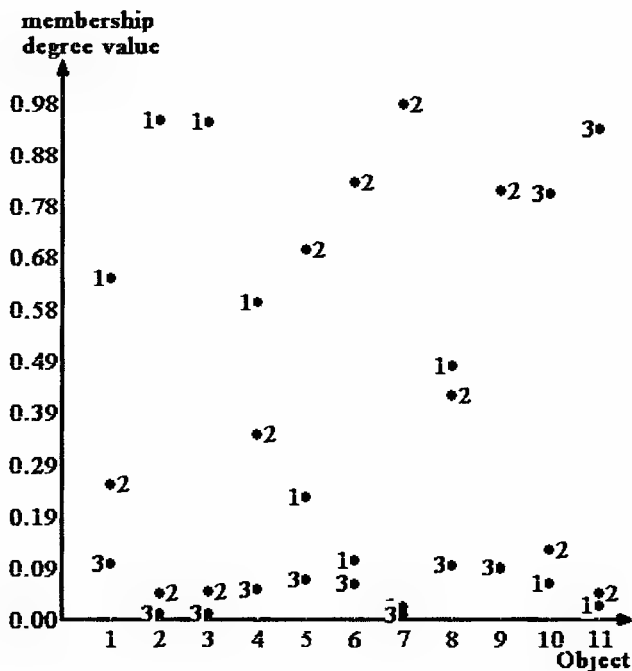


Рис. 4.7. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Педрича на три класса

Результаты разбиения исследуемой совокупности объектов алгоритмом Педрича на четыре класса приведены в таблице 4.13, а линейная диаграмма представлена на рис. 4.8. В качестве помеченных выступали объекты x_3 , x_6 , x_9 и x_{11} с соответствующими функциями принадлежности $\mu_{13} = 1$, $\mu_{26} = 1$, $\mu_{39} = 1$ и $\mu_{411} = 1$.

Таблица 4.13

Значения принадлежностей объектов классам при разбиении исследуемой совокупности алгоритмом Педрича на четыре класса

Номер класса	Объект										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.538	0.897	0.944	0.392	0.051	0.007	0.102	0.270	0.045	0.068	0.015
2	0.241	0.039	0.029	0.160	0.052	0.980	0.248	0.140	0.168	0.135	0.038
3	0.146	0.054	0.022	0.402	0.878	0.009	0.619	0.521	0.739	0.122	0.025
4	0.075	0.010	0.005	0.046	0.019	0.004	0.031	0.069	0.048	0.675	0.922

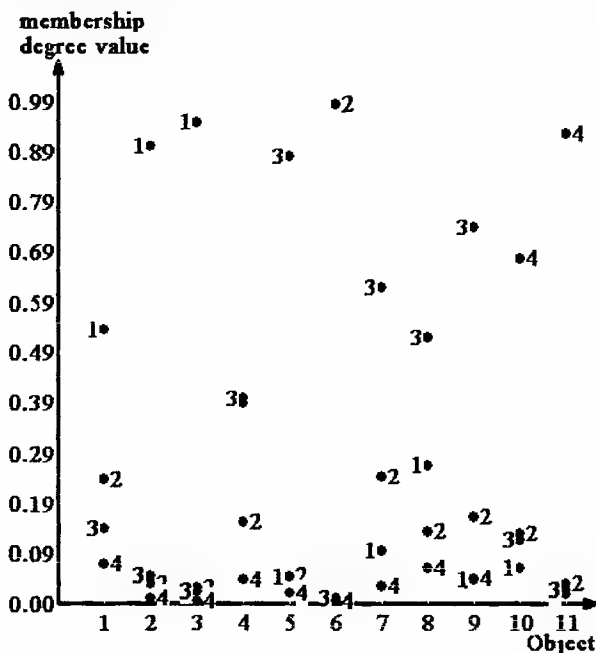


Рис. 4.8. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Педрича на четыре класса

При обработке данных алгоритмом Беждека — Данных объекты x_{10} и x_{11} всегда попадают в один класс с высоким значением принадлежности: при разбиении на два и четыре класса они образуют ядро группы с номером 1, при разбиении на три класса — ядро группы номер 2. Объект x_9 занимает промежуточное положение между хорошо выделяющимися группами $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ и $\{x_{10}, x_{11}\}$, что особенно заметно при разбиении на два класса, а при разбиении на четыре класса объект x_9 выделяется в класс номер 2.

Результаты обработки данных алгоритмом Педрича отличаются от результатов, полученных при помощи алгоритма Беждека — Данных, не только значениями принадлежности объектов классам, но и распределением объектов по классам: к примеру, при разбиении на два класса ядра групп выглядят как $\{x_1, x_2, x_3, x_4, x_5, x_7, x_8\}$ и $\{x_6, x_9, x_{10}, x_{11}\}$. При разбиении на три и четыре класса результаты работы обоих алгоритмов также довольно существенно отличаются друг от друга, что объясняется зависимостью результатов разбиения от элементов множества помеченных объектов при использовании алгоритма Педрича.

В таблице 4.14 приведены результаты разбиения исследуемой совокупности объектов алгоритмом Уиндхема на два класса.

Таблица 4.14

Результаты разбиения исследуемой совокупности объектов алгоритмом Уиндхема на два класса

Объект	Принадлежности объектов		Веса прототипов	
	Номера кластеров		Номера кластеров	
	1	2	1	2
x_1	0.5018	0.4982	0.0917	0.0914
x_2	0.5020	0.4980	0.0895	0.0891
x_3	0.5021	0.4979	0.0952	0.0948
x_4	0.5018	0.4982	0.1012	0.1009
x_5	0.4998	0.5002	0.1108	0.1109
x_6	0.4999	0.5001	0.1082	0.1083
x_7	0.5002	0.4998	0.1177	0.1178
x_8	0.5001	0.4999	0.1043	0.1044
x_9	0.4988	0.5012	0.0862	0.0867
x_{10}	0.4990	0.5010	0.0522	0.0524
x_{11}	0.4991	0.5009	0.0430	0.0433

Соответствующая линейная диаграмма представлена на рис. 4.9.

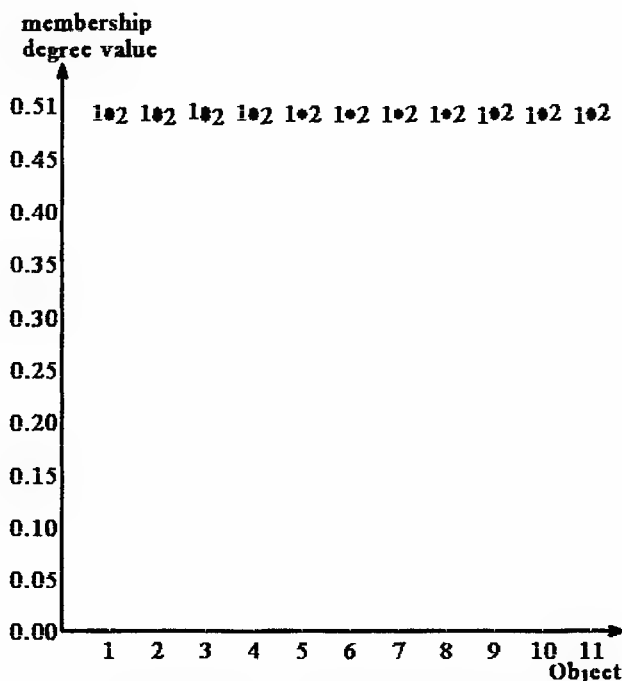


Рис. 4.9. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Уиндхема на два класса

Результаты, полученные при обработке исследуемой совокупности объектов алгоритмом Уиндхема, достаточно интересны и заслуживают более детального рассмотрения. В первую очередь следует отметить весьма небольшие отличия значений степеней принадлежности обоим классам, однако анализ матрицы разбиения позволяет выделить группы $\{x_1, x_2, x_3, x_4, x_7, x_8\}$ и $\{x_5, x_6, x_9, x_{10}, x_{11}\}$, которые, принимая во внимание только матрицу нечеткого разбиения, довольно плохо различаются между собой, что весьма иллюстративно подтверждается линейной диаграммой. Еще одним интересным обстоятельством является то, что центром первого класса оказывается объект x_3 , а объ-

ект x_9 может быть выделен как центр второго класса. Эти результаты, если не принимать во внимание значения степеней принадлежности объектов классам, почти полностью совпадают с результатами, полученными при обработке данных алгоритмом Педрича: исключение составляет объект x_5 , перемещенный при обработке данных алгоритмом Уиндхема во вторую группу. В то же время анализ результатов обработки данных, полученных при помощи алгоритма Беждека — Данна, позволяет выделить центрами соответствующих групп объекты x_7 и x_{10} .

Весьма интересны также результаты, представленные матрицей весов прототипов классов. Отчетливо выделяется группа $\{x_4, x_5, x_6, x_7, x_8\}$, для элементов которой имеет место $v_{ii} > 0.1000$. Если данную группу интерпретировать как «ядро» всей исследуемой совокупности, то, очевидно, центром исследуемого множества объектов следует считать объект x_7 . Вместе с тем, для объектов x_3 и x_9 значение «степени прототипности» $v_{ii} < 0.1000$.

Таблица 4.15 представляет результаты обработки исследуемой совокупности объектов алгоритмом Уиндхема при разбиении на три класса.

Таблица 4.15

Результаты разбиения исследуемой совокупности объектов алгоритмом Уиндхема на три класса

Объект	Принадлежности объектов			Веса прототипов		
	Номера кластеров			Номера кластеров		
	1	2	3	1	2	3
x_1	0.3344	0.3337	0.3319	0.0917	0.0916	0.0913
x_2	0.3346	0.3337	0.3317	0.0894	0.0893	0.0890
x_3	0.3346	0.3337	0.3317	0.0952	0.0950	0.0947
x_4	0.3345	0.3337	0.3318	0.1012	0.1011	0.1008
x_5	0.3333	0.3333	0.3334	0.1108	0.1108	0.1109
x_6	0.3333	0.3333	0.3334	0.1082	0.1083	0.1084
x_7	0.3334	0.3334	0.3332	0.1178	0.1178	0.1178
x_8	0.3334	0.3334	0.3332	0.1043	0.1043	0.1044
x_9	0.3326	0.3330	0.3343	0.0862	0.0863	0.0867
x_{10}	0.3326	0.3332	0.3342	0.0522	0.0523	0.0526
x_{11}	0.3327	0.3332	0.3341	0.0430	0.0432	0.0434

Линейная диаграмма, иллюстрирующая разбиение исследуемой совокупности объектов алгоритмом Уиндхема на три класса, изображена на рис. 4.10.

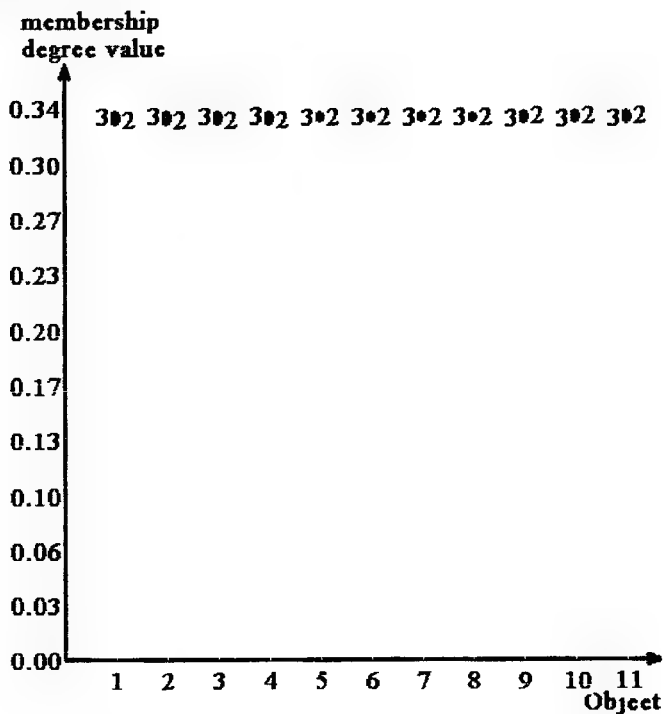


Рис. 4.10. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Уиндхема на три класса

При разбиении исследуемой совокупности объектов алгоритмом Уиндхема на три класса для всех объектов значения степеней принадлежности также различаются между собой весьма незначительно: $\mu_{il} \approx 0.3300, l=1,2,3, i=1, \dots, 11$, — однако анализ матрицы разбиения позволяет выделить группы $\{x_1, x_2, x_3, x_4\}$ и $\{x_9, x_{10}, x_{11}\}$. В то же время объекты группы $\{x_5, x_6, x_7, x_8\}$ распределены по всем трем классам равномерно, что заметно на линейной диаграмме. Следует также от-

метить то обстоятельство, что значения степеней принадлежности объектов x_2 и x_3 всем трем классам совпадают, что, с одной стороны, позволяет любой из этих объектов в равной мере рассматривать в качестве центра первой группы, с другой стороны, однако для объекта x_3 «степень прототипности» выше, чем для объекта x_2 , так что центром первого класса вновь оказывается объект x_3 . В свою очередь, объект x_9 отчетливо выделяется в качестве центра третьего класса, тогда как при обработке данных как алгоритмом Беждека — Данна, так и при их обработке алгоритмом Педрича объект x_9 явно относится к другой группе, чем объекты x_{10} и x_{11} .

Анализ матрицы весов прототипов в данном случае оказывается малоприменимым для интерпретации полученных результатов, поскольку, аналогично разбиению на два класса, выделяется группа $\{x_4, x_5, x_6, x_7, x_8\}$, для элементов которой имеет место $v_{ii} > 0.1000$, однако только объект x_4 может быть отнесен к первой группе в силу значения степени принадлежности.

Результаты обработки исследуемой совокупности объектов алгоритмом Уиндхема при числе классов, равном четырем, представлены в таблице 4.16.

Таблица 4.16

Результаты разбиения исследуемой совокупности объектов алгоритмом Уиндхема на четыре класса

Объект	Принадлежности объектов				Веса прототипов			
	Номера кластеров				Номера кластеров			
	1	2	3	4	1	2	3	4
x_1	0.2515	0.2501	0.2503	0.2481	0.0918	0.0915	0.0916	0.0912
x_2	0.2517	0.2501	0.2503	0.2479	0.0896	0.0893	0.0893	0.0889
x_3	0.2518	0.2501	0.2503	0.2478	0.0953	0.0950	0.0950	0.0945
x_4	0.2516	0.2501	0.2503	0.2480	0.1014	0.1011	0.1011	0.1006
x_5	0.2498	0.2501	0.2500	0.2501	0.1107	0.1108	0.1108	0.1110
x_6	0.2499	0.2501	0.2500	0.2500	0.1082	0.1083	0.1083	0.1084
x_7	0.2501	0.2501	0.2501	0.2497	0.1177	0.1178	0.1178	0.1178
x_8	0.2501	0.2501	0.2500	0.2498	0.1043	0.1043	0.1043	0.1044
x_9	0.2489	0.2499	0.2498	0.2514	0.0860	0.0864	0.0863	0.0869
x_{10}	0.2491	0.2499	0.2499	0.2511	0.0520	0.0523	0.0523	0.0527
x_{11}	0.2492	0.2499	0.2499	0.2510	0.0430	0.0432	0.0432	0.0436

Линейная диаграмма, иллюстрирующая матрицу разбиения исследуемой совокупности объектов алгоритмом Уиндхема на четыре класса, изображена на рис. 4.11.

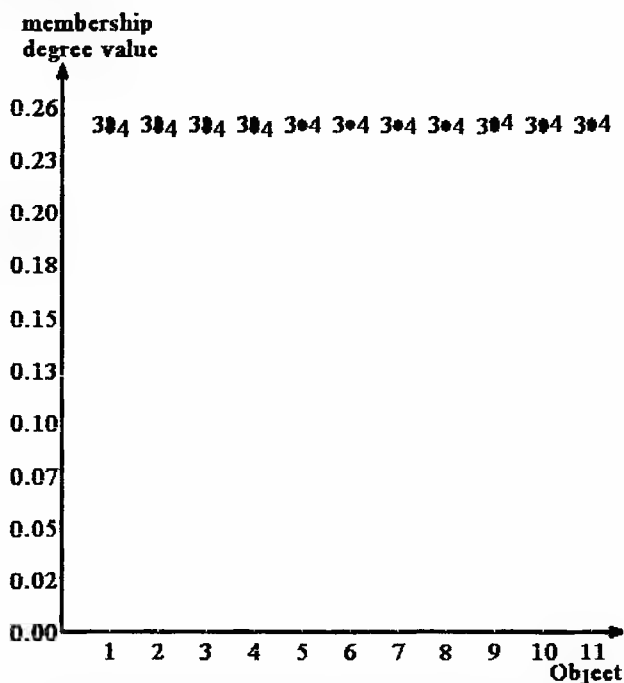


Рис. 4.11. Линейная диаграмма нечеткого разбиения исследуемой совокупности объектов алгоритмом Уиндхема на четыре класса

Результаты разбиения исследуемой совокупности объектов алгоритмом Уиндхема на четыре класса сходны с результатами разбиения на три класса: $\mu_{ii} \approx 0.2500, i=1, \dots, 4, i=1, \dots, 11$, — однако оказывается возможным выделить группы $\{x_1, x_2, x_3, x_4\}$ и $\{x_9, x_{10}, x_{11}\}$, тогда как объекты группы $\{x_5, x_6, x_7, x_8\}$ распределены по всем классам равномерно. Центром первого класса оказывается объект x_3 , а объект x_9 является центром четвертого класса. При рассмотрении матрицы весов прототипов классов так же, как и при разбиениях на два и три

класса, выделяется группа объектов $\{x_4, x_5, x_6, x_7, x_8\}$, для которых $v_{ii} > 0.1000$.

Таким образом, при обработке рассматриваемых исходных данных алгоритмом Уиндхема наиболее просто интерпретируются результаты разбиения на два класса.

Следует указать, что при обработке данных Е. Андерсона алгоритмом Уиндхема отчетливо выделяется класс SETOSA, тогда как классы VERSICOLOR и VIRGINICA «сливаются» [186, с.167-171], что также демонстрировалось Дж. Беждеком при обработке данных Е. Андерсона классическим методом c -средних, методом нечетких c -средних и методами расщепления смесей [53]. Результаты обработки данных Е. Андерсона методом нечетких c -средних при различных значениях показателя нечеткости классификации γ также подробно рассматриваются в работе В. Педрича [144, с.138-139].

В таблице 4.17 представлены значения коэффициента разбиения $F_c(P)$ и энтропии разбиения $H_c(P)$ при разбиении исследуемой совокупности объектов нечеткими оптимизационными кластер-процедурами на два, три и четыре класса.

Таблица 4.17

Значения показателей качества разбиения при обработке исследуемой совокупности объектов различными оптимизационными алгоритмами

Алгоритм	Число классов	Значение	
		коэффициента разбиения $F_c(P)$	энтропии разбиения $H_c(P)$
Алгоритм Беждека — Данна	$c = 2$	0.886705	0.211502
	$c = 3$	0.800439	0.382771
	$c = 4$	0.771432	0.458396
Алгоритм Педрича	$c = 2$	0.686370	0.483679
	$c = 3$	0.687571	0.561724
	$c = 4$	0.627833	0.718427
Алгоритм Уиндхема	$c = 2$	0.500003	0.693144
	$c = 3$	0.333353	1.098610
	$c = 4$	0.250008	1.386290

При рассмотрении значений коэффициента разбиения $F_c(P)$ для различного числа классов уместно указать, что при обработке данных алгоритмом Педрича значения коэффициента разбиения $F_c(P)$ для $c = 2$ и $c = 3$ практически совпадают, тогда как при их обработке методом нечетких c -средних наблюдается существенный «скачок» зна-

чений $F_c(P)$. Сравнивая «скачки» значений обоих показателей разбиения при переходе от $c=2$ к $c=3$ и при переходе от $c=3$ к $c=4$, можно сделать вывод, что, хотя $c=2$ вполне может соответствовать «реальному» числу классов, $c=3$ является более «приемлемым» числом классов в нечетком разбиении, чем $c=2$, что, в свою очередь, соответствует числу классов в экспертном разбиении.

Оценка сходимости нечетких оптимизационных кластер-процедур производилась в соответствии с формулой (4.1) также для случаев разбиения на два, три и четыре класса. Результаты исследования сходимости алгоритма Бежdeka — Данна представлены на рис. 4.12.

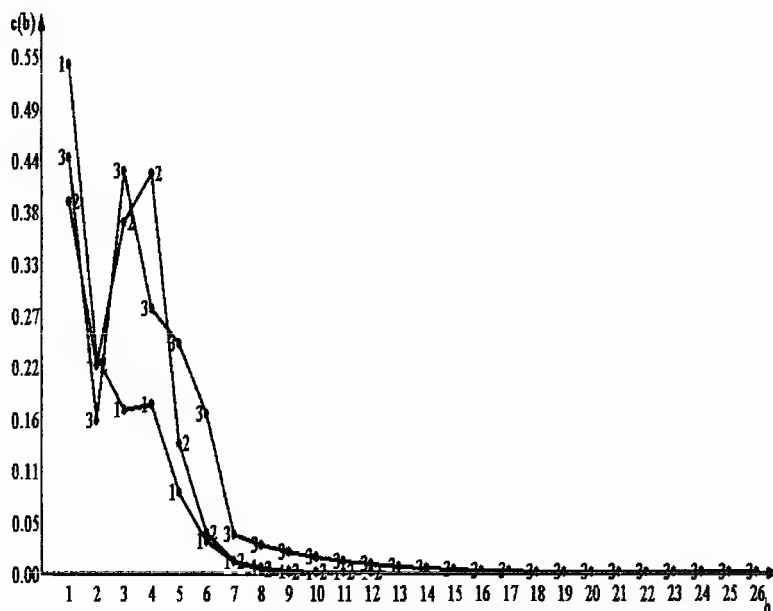


Рис. 4.12. Сходимость алгоритма Бежdeka — Данна при разбиении исследуемой совокупности объектов на два, три и четыре класса

На рис. 4.13 представлены результаты исследования сходимости алгоритма Педрича.

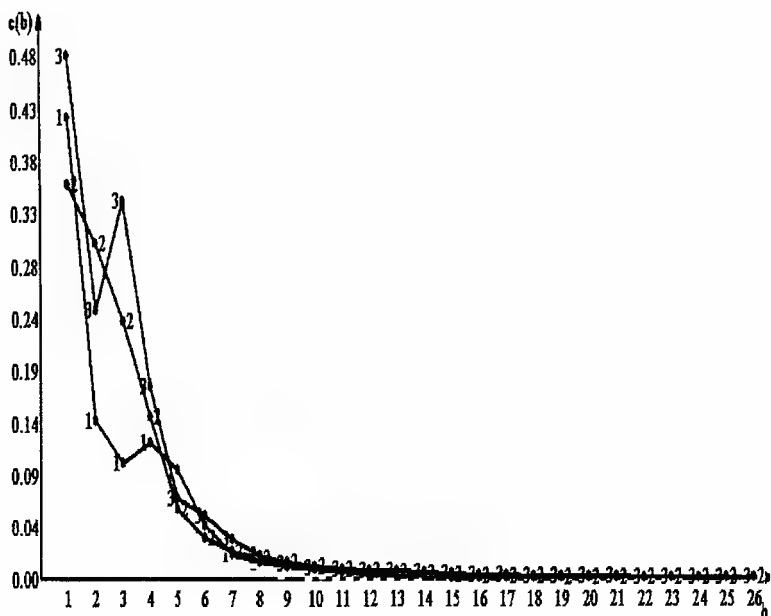


Рис. 4.13. Сходимость алгоритма Педрича при разбиении исследуемой совокупности объектов на два, три и четыре класса

При рассмотрении графиков сходимости следует отметить «скачки» как при работе процедуры Беждека — Данна, так и при работе процедуры Педрича, что также отмечалось В. Педричем [140, с.17-18], причем при рассмотрении графика сходимости процедуры Беждека — Данна существенные «скачки» наблюдаются у графиков под номерами 2 и 3, а при рассмотрении графика сходимости процедуры Педрича — у графика номер 3. Данное обстоятельство подтверждает упоминавшийся выше тезис У. Кеймака и М. Сетнеса о зависимости результатов работы оптимизационных алгоритмов от началь-

го разбиения и целесообразности проведения ряда экспериментов с различными начальными разбиениями.

Иллюстрация сходимости алгоритма Уиндхема при разбиении исследуемой совокупности объектов на два, три и четыре класса представлена на рис. 4.14.

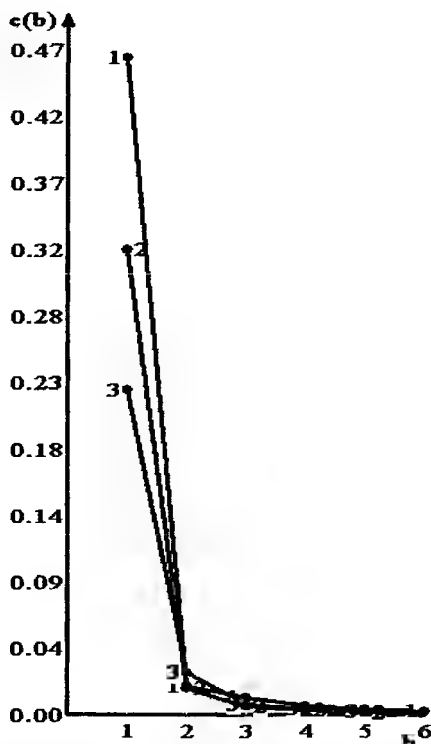


Рис. 4.14. Сходимость алгоритма Уиндхема при разбиении исследуемой совокупности объектов на два, три и четыре класса

На рис. 4.12 — 4.14 график сходимости того или иного метода при разбиении на два класса отмечен цифрой 1, при разбиении на три класса — цифрой 2, при разбиении на четыре класса — цифрой 3.

Рассмотрение представленных результатов работы нечетких оптимизационных методов кластеризации показывают, что наиболее быстрым является алгоритм Уиндхема. Касательно сходимости оптимизационных процедур следует также отметить, что с целью уменьшения числа итераций в тех случаях, когда решаемая задача не требует высокой точности вычислений, удовлетворительные результаты могут быть получены при значении порога ε , приблизительно равном 0.01.

На рис. 4.15 представлены результаты работы иерархического варианта алгоритма Тамуры — Хигути — Танаки.

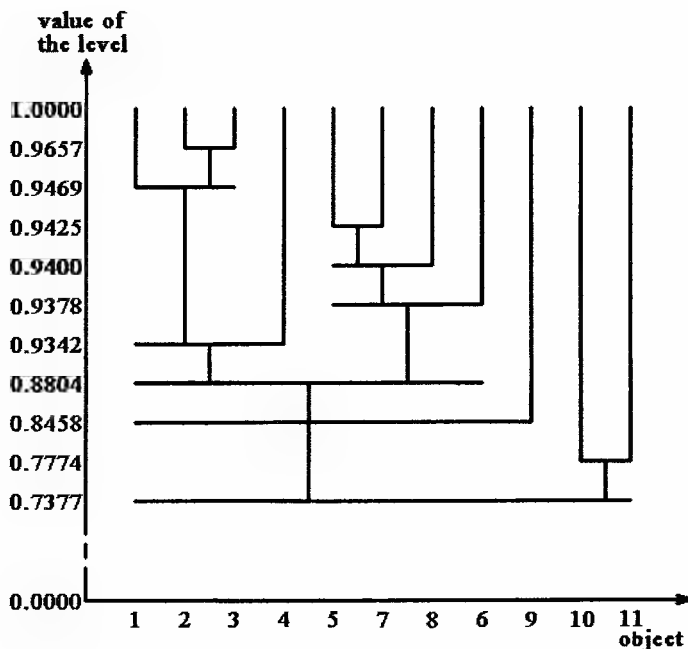


Рис. 4.15. Результаты работы иерархической версии алгоритма Тамуры — Хигути — Танаки

При рассмотрении иерархии в первую очередь необходимо отметить разделение объектов x_{10} и x_{11} при $\alpha > 0.7774$ и выделение объекта x_9 в отдельный класс при $\alpha > 0.8458$. Разбиение, в наибольшей степени соответствующее экспертному, достигается при

$0.8804 < \alpha < 0.9342$ и выглядит следующим образом: первый класс представляет собой множество объектов $\{x_1, x_2, x_3, x_4\}$, второй класс представляет собой множество объектов $\{x_5, x_6, x_7, x_8\}$, объект x_9 представляет собой третий класс, а объекты x_{10} и x_{11} — соответственно четвертый и пятый классы, так что при объединении третьего, четвертого и пятого классов результаты полностью совпадут с экспертным разбиением. Данное обстоятельство указывает на то, что операция (max-min)-транзитивного замыкания в некоторой степени искажает геометрическую структуру исследуемой совокупности объектов, что накладывает ограничения на использование алгоритма Тамуры — Хигути — Танаки при классификации объектов.

На рис. 4.16 изображен исходный граф, иллюстрирующий классифицируемую с помощью алгоритма Берштейна — Дзюбы совокупность объектов.

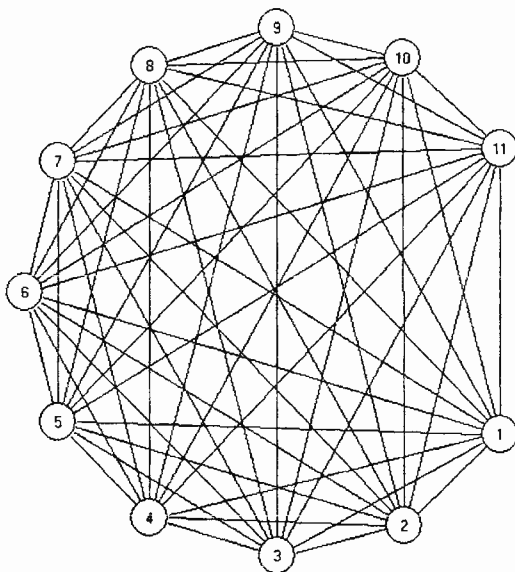


Рис. 4.16. Граф сходства исследуемой совокупности объектов

Результаты работы процедуры классификации на нечетких графах Берштейна — Дзюбы для двух и трех классов представлены на рис. 4.17.

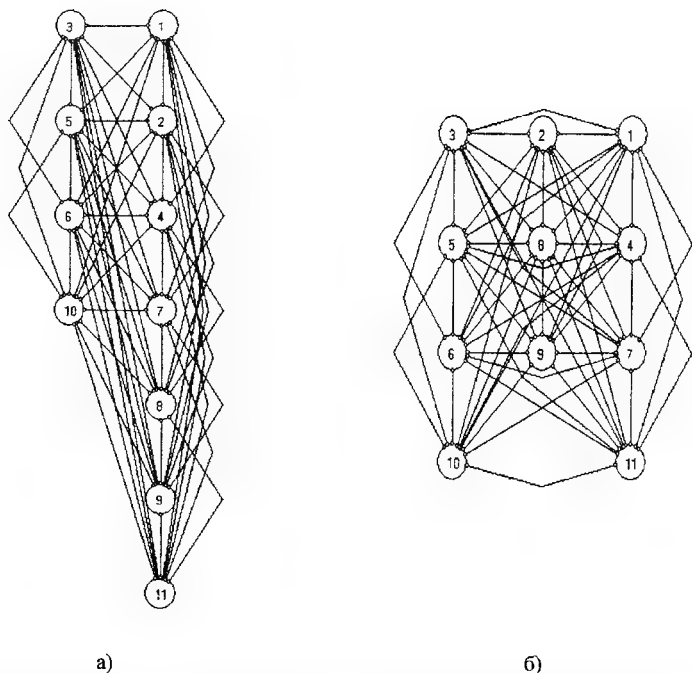


Рис. 4.17. Результаты обработки исследуемой совокупности объектов алгоритмом Берштейна — Дзюбы: а) при числе классов, равном двум; б) при числе классов, равном трем

Очевидно, что процедура классификации на нечетких графах Берштейна — Дзюбы относит наиболее похожие объекты к различным классам, а наиболее различающиеся объекты оказываются в одном классе. Данное обстоятельство позволяет рассматривать алгоритм классификации на нечетких графах Берштейна — Дзюбы как метод решения задачи обратной классификации в рамках концепции двойственности, подробно рассмотренной И. Д. Манделем [31, с.132-136].

Представленные результаты работы различных нечетких кластер-процедур имеют иллюстративный характер, и, безусловно, не могут служить основанием для предпочтения каких-либо одних кластер-процедур перед другими.

Резюме

Цель сравнительного анализа обуславливается либо теоретическими рассмотрениями, либо конкретным прикладным исследованием. В первом случае сравнительный анализ проводится с целью определения эффективности кластер-процедур для решения некоторого класса задач. Во втором случае, как правило, между собой сравниваются результаты работы нескольких алгоритмов для выявления реальной структуры исследуемой совокупности и последующего уточнения результатов. Эффективность любой кластер-процедуры может оцениваться по уровню устойчивости к возмущениям, оценке трудоемкости, а также оценке сходимости алгоритма. Для оценки результатов классификации в случае применения оптимизационных методов могут использоваться различные показатели, в частности, коэффициент разбиения $F_c(P)$, энтропия разбиения $H_c(P)$ и пропорциональная экспонента $E_c(P)$, а также (D, F) -диаграмма, предложенная Е. Траурвертом. Для представления результатов классификации при использовании оптимизационных и эвристических нечетких кластер-процедур следует рассматривать матрицу нечеткого распределения объектов по классам, а при использовании процедуры Кутюрье — Фьолео в распоряжение исследователя должна предоставляться матрица пересечений нечетких классов. В ряде случаев оказывается целесообразным построение графика сходимости процедуры. В распоряжение исследователя необходимо предоставлять также визуальное представление результатов полученной классификации, а в качестве подобного средства может быть использована линейная диаграмма. При выборе в качестве инструмента исследования алгоритма классификации на нечетких графах Берштейна — Дзюбы результатом работы процедуры будет представление нечетких двудольных частей графа, а при использовании иерархических методов нечеткого подхода к решению задачи автоматической классификации в качестве результатов классификации исследователю должны предоставляться дендрограмма или дерево нечеткой иерархии.

МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ НЕЧЕТКОГО ПОДХОДА К РЕШЕНИЮ ЗАДАЧИ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

5.1. Общая схема применения нечетких методов решения задачи автоматической классификации

5.1.1. Выбор типа метода нечеткого подхода к решению задачи автоматической классификации

В рамках прикладного исследования, в случае выбора кластерного анализа в качестве инструмента решения проблемы классификации, возникает вопрос о выборе метода решения конкретной задачи, что осложняется достаточно большим числом кластер-процедур. Данная проблема детально рассматривается И. Д. Манделем [31, с.148-156], а также в работе [77] и некоторых других исследованиях. Несмотря на незначительное, по отношению к остальным методам, число нечетких кластер-процедур, выбор наиболее адекватного метода нечеткой кластеризации в конкретном случае также может представлять собой довольно серьезную проблему. Общая схема выбора нечеткой кластер-процедуры, предложенная в работе [18], предусматривает два этапа: обоснование выбора одного из трех рассматриваемых типов методов нечеткого подхода к кластеризации и выбор конкретной кластер-процедуры. Рекомендации по выбору типа метода нечеткого подхода к решению задачи автоматической классификации можно выработать, исходя из целей классификации и имеющихся содержательных соображений о компактности выделяемых групп. Эти рекомендации могут быть сформулированы следующим образом:

- 1) если у исследователя существуют содержательные представления об условиях объединения объектов в классы, следует выбрать группу эвристических методов нечеткого подхода в кластерном анализе;
- 2) если целью классификации является получение нечеткого разбиения на заранее известное число классов исследуемой совокупности объектов, то следует выбрать группу оптимизационных методов нечеткого подхода в кластерном анализе;

3) если целью классификации является получение наглядного представления о нечеткой структуре классифицируемой совокупности объектов сравнительно небольшого объема, то следует выбрать иерархические методы нечеткого подхода в кластерном анализе.

Безусловно, при определении типа метода необходимо учитывать априорную информацию, к примеру, о возможных параметрах кластер-процедуры, а также формальные представления о классификации. Таким образом, изложенные правила выбора направления нечеткого подхода к решению задачи автоматической классификации могут быть представлены в виде схемы, изображенной на рис. 5.1.

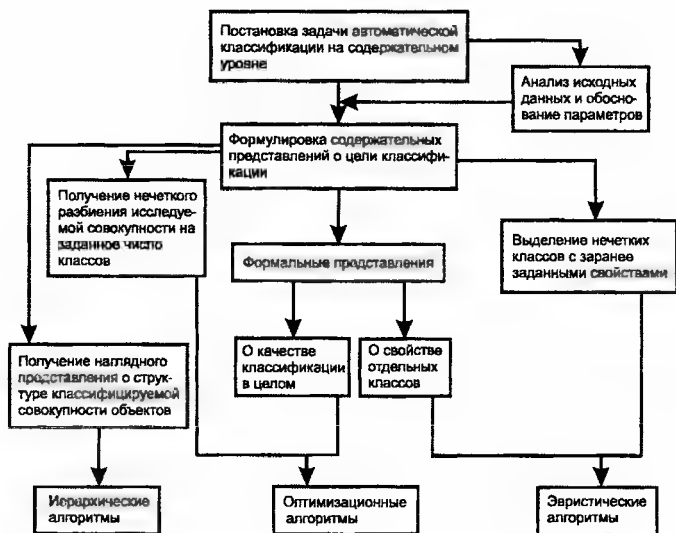


Рис. 5.1. Схема выбора типа метода нечеткого подхода в кластер-анализе

Следует отметить, что при обосновании типа метода нечеткого подхода к решению задачи автоматической классификации все этапы такого обоснования целесообразно проводить именно в том порядке, в котором они приведены на схеме.

5.1.2. Выбор алгоритма нечеткого подхода к решению задачи автоматической классификации и обоснование параметров

После того как определен тип метода нечеткого подхода к решению задачи автоматической классификации, исходя из вида матрицы исходных данных и имеющейся априорной информации, следует выбрать алгоритм, адекватный условиям конкретной задачи.

Если для решения задачи классификации выбрано эвристическое направление нечеткого подхода, то главным критерием выбора алгоритма является соответствие особенностей того или иного алгоритма содержательной постановке задачи. Специфика нечетких методов автоматической классификации эвристического направления отражена в таблице 5.1.

Таблица 5.1

Характеристики эвристических методов нечеткого подхода в кластерном анализе

Метод	Вид матрицы исходных данных	Параметры алгоритма	Примечания
Метод Гитмана — Левина	$X_{\text{исх}} = [x_i^t]$	$\hat{d}$	Алгоритм использует как значение функции $\mu_B(x)$, так и метрику d
Метод Тамуры — Хигути — Танаки	$r_{\text{исх}} = [\mu_T(x_i, x_j)]$	c	Алгоритм использует транзитивное замыкание исходной матрицы сходства объектов
Метод Купюрье — Фюлсео	$d_{\text{исх}} = [\mu_I(x_i, x_j)]$	α, u, w	Алгоритм допускает пересечение нечетких кластеров
Метод Берштейна — Дзюбы	$r_{\text{исх}} = [\mu_T(x_i, x_j)]$	отсутствуют	Алгоритм выделения максимальной двудольной части в нечетком графе

Как указывалось ранее, алгоритм Гитмана — Левина позволяет быстро обрабатывать достаточно большие массивы данных, а полученные в результате кластеры могут иметь как эллиптическую, так и сферическую форму. Таким образом, эта процедура может рассматриваться как процедура комбинированной прямой классификации в том смысле, в котором этот термин использовался И. Д. Манделем [31, с.37]. Результаты применения процедуры Тамуры — Хигути — Танаки, как показали вычислительные эксперименты, могут оказаться не-

корректными, так что данную процедуру следует применять в качестве инструмента предварительного анализа исследуемой совокупности, причем в ее иерархическом варианте. Процедура Берштейна — Дзюбы позволяет выделять кластеры сложной формы, и, как указывалось выше, в принципе возможно выделение большего чем два числа классов, однако данная процедура, как и все процедуры классификации на графах, позволяет классифицировать объекты исследуемой совокупности сравнительно небольшого объема. Алгоритм Кутюрье — Фьолео предлагает исследователю весьма гибкий аппарат для анализа данных и не предполагает жестких ограничений на объем исследуемой совокупности.

Таким образом, конкретный алгоритм эвристического направления нечеткого подхода к решению задачи автоматической классификации может быть выбран в соответствии со следующими рекомендациями:

- 1) если число объектов исследуемой совокупности достаточно велико и на множестве $X = \{x_1, \dots, x_n\}$ оказывается возможным определить нечеткое множество $B = \{(x_i, \mu_B(x_i))\}, i = 1, \dots, n$, то следует обосновать выбор метрики d и порога $\hat{d}$ и использовать алгоритм Гитмана — Левина;
- 2) если целью классификации является предварительный анализ исследуемой совокупности объектов, в процессе которого требуется получить разбиение на c четких классов, то следует использовать алгоритм Тамуры — Хигути — Танаки;
- 3) если допускается пересечение нечетких кластеров, а также имеются предположения о минимальном числе u объектов в каждом нечетком кластере, то следует обосновать выбор порога α и использовать алгоритм Кутюрье — Фьолео;
- 4) если число элементов множества $X = \{x_1, \dots, x_n\}$ сравнительно невелико, а также существуют предположения о сложной форме кластеров и требуется визуальное представление результатов классификации, то следует выбрать алгоритм классификации на нечетких графах Берштейна — Дзюбы.

Если же для решения задачи выбрано оптимизационное направление, главной проблемой оказывается обоснование вида функционала. Поскольку выбор функционала определяется, помимо формы матрицы исходных данных, вида шкалы, в которой измерены признаки, и типа признакового пространства, также спецификой конкретной задачи, то при выборе функционала качества разбиения целесообразно

учитывать содержательную интерпретацию функционала, что подробно рассматривалось выше.

Основные характеристики нечетких методов автоматической классификации оптимизационного направления представлены в таблице 5.2.

Таблица 5.2

Характеристики оптимизационных методов нечеткого подхода в кластерном анализе

Метод (критерий)	Вид матрицы исходных данных	Параметры алгоритма	Примечания
Метод Распина ($Q_1^I(P)$)	$d_{\text{расп}} = [\mu_i(x_i, x_j)]$	$c, \hat{d}$	Оптимизирует разницу между контекстной и прямой близостью объектов
Метод Уиндхема ($Q_2^I(P)$)	$d_{\text{расп}} = [\mu_i(x_i, x_j)]$	c	Оптимальное разбиение достигается при минимальном отклонении нечетких принадлежностей объектов от нечетких центров кластеров
Метод Рубенса ($Q_3^I(P)$)	$d_{\text{расп}} = [\mu_i(x_i, x_j)]$	c	Оптимальное разбиение достигается при минимальном взаимном отклонении элементов в каждом нечетком кластере
Метод Даве — Сена ($Q_5^I(P)$)	$d_{\text{расп}} = [\mu_i(x_i, x_j)]$	c, γ	Функционал представляет собой обобщение критерия Кофмана — Рауссеу
Метод Беждека — Данна ($Q_1^{II}(P)$)	$X_{\text{расп}} = [x'_i]$	c, γ	Оптимальное разбиение достигается при минимальных нечетких отклонениях объектов от центров нечетких кластеров
Метод Педрича ($Q_2^{II}(P)$)	$X_{\text{расп}} = [x'_i]$	c	Метод использует частично обучающую выборку
Метод Райга ($Q_3^{II}(P)$)	$X_{\text{расп}} = [x'_i]$	c	Функционал представляет собой взвешенный аналог корреляционного отношения

Вместе с тем, в случае, когда выявление специфики задачи вызывает некоторые трудности, целесообразно использовать, как это указывалось при рассмотрении схемы сравнения нечетких кластер-

процедур, несколько функционалов, после чего выбрать наиболее эффективный критерий. Как отмечал в связи с этим И. Д. Мандель, «самый надежный способ обоснования вида критерия — это придание ему четкого содержательного смысла, совпадающего с конечным результатом, найденным в данной задаче или максимально близким к нему» [31, с.153]. В случае, когда выбранный функционал является одним из функционалов семейства $Q''(P)$, при отсутствии предположений о форме и взаимном расположении кластеров, может возникнуть проблема выбора расстояния. В этом случае целесообразно проводить исследование с использованием различных метрик, как это рекомендуется С. А. Айвазяном [37, с.328]. Вместе с тем, для выбора алгоритма оптимизационного направления также могут быть сформулированы правила, подобные рекомендациям для выбора эвристической нечеткой кластер-процедуры.

Касательно процедур иерархического направления следует отметить, что, поскольку данные процедуры не предусматривают задания входных параметров, то при выборе процедуры достаточно руководствоваться видом матрицы исходных данных и спецификой решаемой задачи. Особенности рассмотренных иерархических процедур представлены в таблице 5.3.

Таблица 5.3

Характеристики иерархических методов нечеткого подхода в кластерном

анализе

Метод	Вид матрицы исходных данных	Параметры алгоритма	Примечания
Метод Ватады — Танаки — Асаи	$r_{\text{пол}} = [\mu_T(x_i, x_j)]$	отсутствуют	Возможны восходящая и нисходящая версии
Метод Думитреску	$X_{\text{пол}} = [x'_i]$	отсутствуют	Бинарный дивизимный алгоритм

Правила выбора нечеткой кластер-процедуры иерархического направления также весьма просты:

- 1) если исходные данные заданы матрицей $X_{\text{пол}} = [x'_i]$, то следует использовать бинарный дивизимный алгоритм Думитреску;
- 2) если исходные данные заданы матрицей $r_{\text{пол}} = [\mu_T(x_i, x_j)]$, то следует использовать эвристический алгоритм Ватады — Танаки — Асаи.

Следует, однако, отметить, что если исходные данные заданы в виде матрицы «объект — свойство», но в силу специфических осо-

бенностей задачи или содержательных соображений необходимо прибегнуть к восходящей стратегии классификации, то следует перейти к матрице близости объектов «объект — объект» и воспользоваться восходящей версией алгоритма Ватады — Танаки — Асаи.

Безусловно, предложенная схема выбора типа метода и алгоритма не претендует на универсальность и носит общий характер. Вместе с тем, следование рекомендациям предлагаемой методологии позволит исследователю, не знакомому со спецификой нечетких методов автоматической классификации, быстрее сориентироваться среди алгоритмов нечеткого подхода в кластерном анализе с целью выбора наиболее адекватного для решения конкретной задачи метода. Более того, рекомендации предложенной методологии могут послужить основой для разработки консультационной экспертной системы выбора конкретного алгоритма.

Касательно проблемы обоснования параметров следует отметить, что данный вопрос, в общем, рассматривался И. Д. Манделем [31, с.156-159], так что в данном случае целесообразно указать лишь специфические особенности, присущие данной проблеме при использовании для решения задачи автоматической классификации нечетких методов.

Число классов c требуется задавать в качестве входного параметра в алгоритме Тамуры — Хигути — Танаки, а также во всех оптимизационных кластер-процедурах и зачастую может быть определено на основании сущности решаемой задачи. Если же число классов априори неизвестно, то целесообразно задать ряд значений $c_*, \leq c \leq c^*$, где c_* — наименее возможное число классов, $c_* \geq 2$, c^* — наиболее возможное число классов, $c^* \leq n$, после чего, при использовании алгоритма Тамуры — Хигути — Танаки, выбрать наиболее приемлемое разбиение, исходя из содержательных рассмотрений. В случае использования оптимизационных методов наилучшее разбиение можно выбрать на основании анализа значений показателей качества разбиения, к примеру, коэффициента разбиения $F_c(P)$ или энтропии разбиения $H_c(P)$. Другим способом определения числа классов перед обработкой данных оптимизационной кластер-процедурой является их первоначальная обработка нечеткой эвристической кластер-процедурой, как это предлагалось в работе И. И. Елисеевой и В. О. Рукавишникова [22, с.53-56], либо нечеткой иерархической кластер-процедурой, что предлагается Д. Думитреску [81]. Подобный подход позволяет, по меньшей мере, значительно сократить интервал

$c_* \leq c \leq c^*$, для которого рекомендуется проведение серии экспериментов. Вообще же, как отмечал И. Д. Мандель, определение числа классов представляет собой «узловую проблему кластер-анализа, и неудивительно, что она не находит однозначного решения» [31, с.158].

Число объектов в классе u в алгоритме Кутюрье — Фьолео определяется спецификой решаемой задачи: к примеру, при формировании армейского подразделения требуется разбить множество $X = \{x_1, \dots, x_n\}$ солдат на отделения; в таком случае, помимо учета уровня физической подготовки, профессиональных и специальных навыков, психологической совместимости кандидатов и других признаков, следует учитывать, что в составе каждого отделения должно быть не менее u человек. В случае, когда жесткие ограничения на минимальное количество объектов в классе отсутствуют, целесообразно полагать $u = 1$.

Порог различия объектов α в алгоритме Кутюрье — Фьолео определяется исследователем исходя из содержательных соображений о компактности формируемых групп.

Число объектов в области пересечения классов w в алгоритме Кутюрье — Фьолео так же, как и параметр u , задается в зависимости от особенностей задачи. При решении конкретной задачи рекомендуется провести серию экспериментов с различными значениями u и w и получением ряда результатов, на основании чего выбрать некоторое лучшее, в содержательном или формальном смысле, решение.

Порог $\hat{d}$ выбирается в зависимости от используемой процедуры: в алгоритме Гитмана — Левина данный порог определяется исследователем в зависимости от вида функции принадлежности $\mu_B(x)$ в каждом конкретном случае, что отмечалось непосредственно при рассмотрении алгоритма, а в алгоритме Распини его можно задавать как среднюю связь в классе [31, с.159].

Показатель нечеткости классификации γ , как уже отмечалось, чаще всего полагается равным 2. В работе [41] его предлагается варьировать в пределах от 2 до 5, а В. Педрич [144, с.134] рекомендует задавать γ в интервале от 1.5 до 30. Очевидно, значение γ должно определяться исследователем исходя из количества объектов исследуемой совокупности и результатов предварительной обработки данных.

5.2. Прикладные аспекты нечеткой классификации

5.2.1. Применение нечетких методов автоматической классификации к обработке изображений

Специфика применения нечеткого подхода к решению задачи автоматической классификации обуславливается, как правило, рамками конкретного исследования. Среди исследований, посвященных рассмотрению данной проблемы, в первую очередь необходимо указать на работы [46], [101], [128]. Весьма интересны также результаты исследований немецкого специалиста Б. Штраубе, в работе [158] рассматривающего методологические аспекты нечеткой кластеризации, а также опыт применения нечеткой кластеризации к решению задач в таких областях, как экология, управление, медицина. В работе [159] Б. Штраубе описывает пакет прикладных программ FDA, реализованный на языке FORTRAN IV для мини-ЭВМ типа PDP-11. Среди областей применения нечетких методов автоматической классификации следует отметить применение алгоритма Педрича к проблеме обработки результатов электрокардиографии [140] и применение алгоритма Беждека — Данна к распознаванию рукописных символов [144]. Некоторые другие, но, конечно, далеко не все области применения нечетких методов автоматической классификации представлены в таблице 5.4.

Таблица 5.4

Обзор публикаций по некоторым приложениям нечетких методов автоматической классификации

Область применения нечетких методов автоматической классификации	Ссылка на источник
Обработка изображений	[102], [168], [169], [121]
Распознавание символов	[144]
Экологические исследования	[133], [158]
Классификация сигналов	[140]
Другие	[158], [71], [144], [128]

Наиболее широкое применение нечеткие методы автоматической классификации получают при решении задач обработки изображений, так что целесообразно кратко рассмотреть особенности их применения именно в этой области. В первую очередь следует указать на работы [102], [121], [168], [169], в которых рассматриваются различные примеры применения алгоритма Беждека — Данна к задаче распознавания изображений. В результате применения метода нечетких s -средних к анализируемому изображению определяется число

экспериментально формируемых классов, после чего применяется нечеткий оператор для выявления локальных изменений принадлежности элементов изображения и построения контура объекта. Подобный оператор позволяет осуществить нечеткое разбиение ячеек изображения на два кластера: края изображения и остальных участков изображения.

В работе [169] рассматривается распознавание изображения, представленного на рис. 5.2, полученного методом телеметрии, с изменением его уровня. На нижнем уровне осуществляется грубая кластеризация. К примеру, на аэрофотоснимке города распознаются в общих чертах торговые центры, жилые кварталы и другие объекты, после чего проверяется однородность, и если обнаруживается неоднородная область, осуществляется её распознавание на более высоком уровне — производится деление на строения, автомобили и другие объекты.

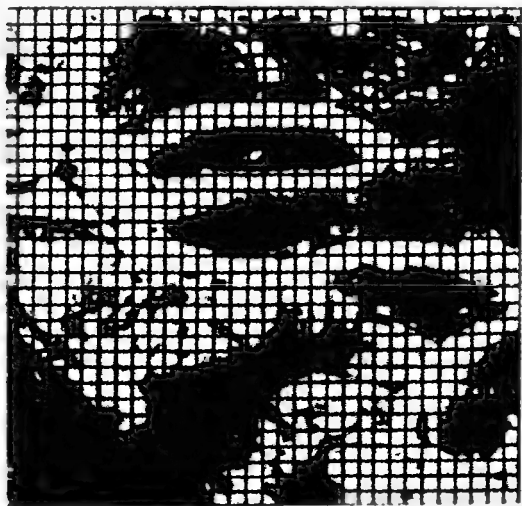


Рис. 5.2. Пример телеметрического изображения для распознавания

В процессе применения алгоритма Беждека — Данна одни и те же участки обводятся с использованием трёх типов спектра. Изображение делится на 32×32 элемента, а каждый элемент, в свою очередь, состоит из 10×10 ячеек. В рассматриваемом случае целесообразно

деление на четыре кластера: река, постройка, земля, лес. Результат, полученный по методу Беждека — Данна, сравнивается с результатом прорисовки границ, сделанной экспертом, после чего, при необходимости, алгоритм применялся вторично с другими значениями входных параметров. При получении удовлетворительного результата осуществляется переход на второй уровень, где первое применение метода проводится для 16×16 элементов, а каждый элемент состоит из 20×20 ячеек. Таким образом, данные имеют пирамидальную структуру, а сам анализ аэрофотоснимка носит иерархический характер. Окончательные результаты обработки изображения с помощью алгоритма Беждека — Данна показали его высокую эффективность.

Вместе с тем, следует указать, что смежные кластеры часто перекрываются, что является причиной некорректной классификации. Вышеописанный подход, предложенный Дж. Беждеком и М. М. Триведи, во-первых, позволяет производить сегментацию только монохромных изображений, а во-вторых, для его успешного применения требуется априорная информация о числе объектов, присутствующих на обрабатываемом снимке и представляющих собой участки, подлежащие сегментации. Этих недостатков лишен подход, предложенный в работе [121] Ю. В. Лимом и С. У. Ли, позволяющий производить сегментацию цветных изображений с автоматическим обнаружением числа распознаваемых областей. Подход, представляющий собой своего рода комбинацию грубой сегментации (coarse segmentation) и сегментации высокого уровня (fine segmentation), состоит в задании порога и применении метода нечетких c -средних.

Основой предложенного метода является предположение о том, что гистограмма представляет собой мультимодальную функцию. На этапе грубой кластеризации производятся обнаружение числа классов и задание порога на основе применения техники фильтрации к гистограммам, так что число «вершин» гистограммы соответствует числу классов, а задание порогов осуществляется на основании обнаружения расположения «впадин» гистограммы. Пикселы, которые не расклассифицированы на первом этапе, распределяются по классам на втором этапе, использующем метод нечетких c -средних. Процесс сегментации представляется авторами подхода в виде следующих процедур:

COARSE SEGMENTATION

DO BEGIN

Получение гистограмм трех цветов изображения;

Применение к гистограммам техники фильтрации;

```

Нахождение пороговых значений;
Обнаружение классов — участков изображения;
FOR  $x := 1$  TO 256
    DO FOR  $y := 1$  TO 256
        DO BEGIN
            IF пиксел  $P(x, y)$  принадлежит какому-
            либо обнаруженному классу
            THEN обозначить этот класс как  $P(x, y)$ 
            ELSE  $P(x, y)$  не классифицирован;
        END
    END
    Вычисление центров  $\tau^l, l = 1, \dots, c$  обнаруженных классов;
END

```

FINE SEGMENTATION

```

DO BEGIN
    FOR  $x := 1$  TO 256
        DO FOR  $y := 1$  TO 256
            DO BEGIN
                Вычисляются значения принадлежности для
                нерасклассифицированных пикселей;
                Выделяются классы, для которых значение
                принадлежности максимально;
            END
        END
    END
    END

```

Сравнение предложенного подхода производилось с тремя другими техниками сегментации и выявило его достаточно хорошие характеристики. Использование грубой сегментации, помимо определения числа классов, позволяет также значительно уменьшить число вычислений на этапе сегментации высокого уровня при применении метода нечетких c -средних.

Более полный обзор различных методов обработки изображений с применением средств нечеткой математики, в том числе нечетких методов автоматической классификации, приводится в работе [41].

5.2.2. Нечеткие методы автоматической классификации и другие нечеткие методы распознавания образов

Кластерный анализ представляет собой методологию распознавания образов, специфика которой заключается в отсутствии обучающих выборок. Вместе с тем, при решении ряда задач могут быть использованы и другие нечеткие подходы к решению задачи распознавания образов. Рассмотрение вопросов, относящихся к нечетким методам кластерного анализа, целесообразно дополнить кратким рассмотрением других методов классификации, имеющих основой теорию нечетких множеств. Подобный обзор является полезным с методологической точки зрения и позволяет более полно рассмотреть современное состояние нечеткого подхода к распознаванию образов и сфер его применения.

Как уже отмечалось, кластерный анализ применяется не только для классификации объектов, но и для решения задачи выделения признаков из множества данных. Различные аспекты применения нечетких методов автоматической классификации к решению этой задачи рассматриваются в работах [56], [81], [104]. Среди других нечетких подходов к решению задачи выделения признаков особо следует отметить методы, предложенные в работах [78], [94], [95], а также подход, предложенный Д. В. Гесу и М. Маккароне [89] и имеющий основой теорию возможностей. Данный подход также подробно рассматривается в работе [38, с.183-186]. Кроме того, в работе [38, с.192-195] рассматривается также применение средств нечеткой математики к распознаванию движущихся предметов.

Нечеткие геометрические признаки формы, позволяющие построить нечеткие методы кластеризации при обработке плоских изображений, предлагаются в работе [26]. Сущность предлагаемого подхода заключается в построении геометрии плоскости, так что оказывается возможным ввести геометрические объекты, которые могут быть использованы для описания элементов изображений на дискретной плоскости, такие как точки, отрезки линий, а также множества точек, не являющихся отрезками линий.

Помимо нечеткой кластеризации, для решения задачи распознавания образов используются также другие подходы, в частности, использующие обучающую выборку. Эти подходы, подробно рассматриваемые В. Педричем [144, с.123-131], условно объединяются в три группы: основанные на вычислении нечетких отношений [127], [154],

основанные на понятии сопоставления образов и основанные на процедурах принятия решений. Целесообразно также отметить лингвистический подход к распознаванию образов, обсуждающийся в работах [25], [87], [122], [130], [131], подход на основе так называемых самораспознающих нечетких классификаторов, предлагаемый С. И. Ватлиным [170], [171], [172], а также работы М. Сугэно [160], [161], посвященные применению теории нечетких интегралов к решению задач распознавания образов. Более того, получил развитие подход к решению задач распознавания образов, основанный на использовании нечетких нейронных сетей [62], [72, с.177-212]. Кроме того, определен интерес представляют также подходы к решению задач распознавания, изложенные в работах [80], [112], [113] и [117].

Нечеткие множества как вспомогательный инструментальный могут использоваться для усовершенствования неразмытых классификаторов. Из методов распознавания образов в условиях нестохастической неопределенности, непосредственно не использующих аппарат нечеткой математики, целесообразно отметить работы Р. С. Михальского [125], [126].

Весьма интересным и многообещающим подходом к решению задач распознавания образов является нечеткий дискриминантный анализ. К примеру, Л. И. Бородкиным [13] предложен метод, в основе которого лежит поиск линейной разделяющей функции, минимизирующей суммарную ошибку распознавания, которая, в свою очередь, оценивает степень сходства нечеткого c -разбиения $P = \{A^1, \dots, A^c\}$ с априорным четким разбиением исходного множества $X = \{x_1, \dots, x_n\}$ на c классов.

Для классификации объектов также оказывается удобным использовать нечеткие отношения типа 2 [33], так что решение задачи классификации представляет собой трехэтапный процесс:

1. На классифицируемой совокупности объектов $X = \{x_1, \dots, x_n\}$ задается нечеткое отношение типа 2 $R = \{\mu_R(x_i, x_j) \mid x_i, x_j \in X\}$, функция принадлежности $\mu_R(x_i, x_j)$ которого трактуется как степень уверенности в том, что объекты x_i и x_j принадлежат одному и тому же классу;
2. Нечеткое отношение R типа 2 аппроксимируется отношением подобия S типа 1 или типа 2;

3. На множестве $X = \{x_1, \dots, x_n\}$ строится иерархия классификации на основе сравнения элементов полученного отношения с разными нечеткими порогами по соответствующим отношениям.

Интерес представляет также решение задачи классификации с использованием так называемых упорядоченных взвешенных усредняющих операторов (ordered weighted averaging operators). К примеру, в исследовании Л. Кунчевой [115] предложена схема классификации, в основе которой лежит объединение нескольких классификаторов и агрегирования их решений. Решения отдельных классификаторов обрабатываются как степени принадлежности, распределенные классификатором по классифицируемым объектам. В общем, весь конгломерат классификаторов выступает в едином качестве, как один наилучший классификатор. В исследовании производится сравнение метода классификации на основе упорядоченных взвешенных усредняющих операторов с экспертными оценками, а также с линейным и логарифмическим методами, на основании которого делается вывод о том, что метод, в основе которого лежит аппарат упорядоченных взвешенных усредняющих операторов, обладает лучшими характеристиками, чем сравниваемые с ним методы. Л. Кунчевой приводятся также иллюстративные примеры применения данного подхода для решения задачи классификации.

Для решения задачи классификации применим также метод нечеткого дерева решений. В работе [20] рассматривается задача определения типовых характеристик клиентов некоторой компании на основе информации, содержащейся в базе данных. Типовые характеристики ориентированы на повышение эффективности обслуживания новых клиентов и выработки целенаправленной маркетинговой политики. Для решения задачи предлагается модифицированный алгоритм построения нечетких деревьев решений на основе частичных информационных оценок. Исходные данные представляют собой атрибуты, принимающие значения на некоторых множествах, причем для ряда атрибутов в исходных данных значения отсутствуют. Предлагаемый алгоритм построения нечетких деревьев решений позволяет обрабатывать атрибуты с частично заданными значениями. При построении дерева решений ключевыми задачами являются выбор атрибута, по которому происходит разбиение дерева, и определение условий ограничения роста дерева. Сущность предлагаемого подхода состоит в выборе атрибута с минимальным значением частичной условной энтропии, а для генерации листьев дерева используется понятие частич-

ной совместной информации. Алгоритм построения нечеткого дерева решений заключается в вычислении для всех атрибутов значений частичной совместной информации и частичной условной энтропии, на основании чего происходит выбор текущего узла и определяется наличие листьев строящегося дерева решений, после чего процедура повторяется для остальных атрибутов. Некоторые другие алгоритмы построения нечетких деревьев решений представлены в работах [69], [118], [181], [191].

В работе [137] рассматривается применение средств нечеткой математики для обработки рентгеновских снимков. Метод, предложенный С. К. Палом, Р. А. Кингом и А. А. Хасимом, общая схема которого изображена на рис. 5.3, прост, нагляден и эффективен, так что представляется необходимым его более подробное рассмотрение.

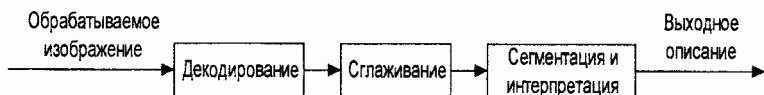


Рис. 5.3. Схема метода Пала — Кинга — Хасима

Сущность подхода, в основе которого лежит алгоритм обнаружения кривых и петель, предложенный Дж. Т. Тоу [165], с последующим сглаживанием и сегментацией, заключается в преобразовании неясных очертаний костных образований на снимке в четкие контуры. Неясные очертания состоят преимущественно из вертикальных, горизонтальных и наклонных линий, так что их можно определить как нечеткие множества:

$$\text{«вертикальные»} = \int_x \mu_v(x) / x;$$

$$\text{«горизонтальные»} = \int_x \mu_h(x) / x;$$

$$\text{«наклонные»} = \int_x \mu_o(x) / x.$$

В свою очередь, функции принадлежности соответствующих нечетких множеств также могут быть определены:

- 1) функция принадлежности вертикальных линий выражается соотношением

$$\mu_v(x) = \begin{cases} 1 - \left| \frac{1}{\operatorname{tg}\theta} \right|^\gamma, & |\operatorname{tg}\theta| > 1; \\ 0, & \text{иначе} \end{cases} \quad (5.1)$$

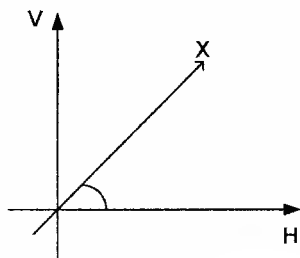
2) функция принадлежности горизонтальных линий определяется соотношением

$$\mu_H(x) = \begin{cases} 1 - |\operatorname{tg}\theta|^\gamma, & |\operatorname{tg}\theta| < 1; \\ 0, & \text{иначе} \end{cases} \quad (5.2)$$

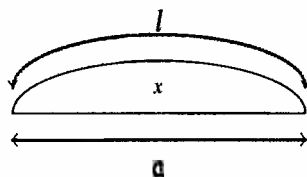
3) функция принадлежности наклонных линий определяется в соответствии с формулой

$$\mu_o(x) = \begin{cases} 1 - \left| \frac{\theta - 45}{45} \right|^\gamma, & 0 < |\operatorname{tg}\theta| < \infty. \\ 0, & \text{иначе} \end{cases} \quad (5.3)$$

В соотношениях (5.1), (5.2), и (5.3) символ γ представляет собой некоторый положительный параметр, с помощью которого подбирается степень нечеткости, а значение величины $\operatorname{tg}\theta$ определяет наклон линии x , что представлено на рис. 5.4 а).



а)



б)

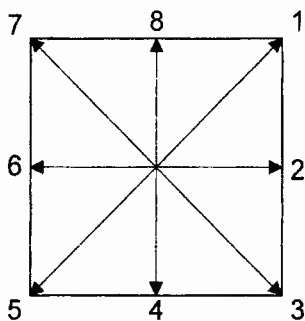
Рис. 5.4. Функции принадлежности: а) линии; б) кривой

В свою очередь, кривизна линии l , соединяющей концы отрезка a некоторого сегмента x , изображенного на рис. 5.4 б), определяется выражением

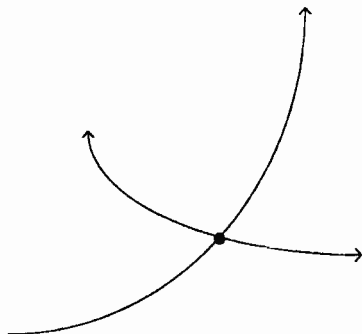
$$\mu_A(x) = \left(1 - \frac{a}{l} \right)^\gamma. \quad (5.4)$$

Очевидно, что с уменьшением величины $\frac{a}{l}$ возрастает кривизна линии l сегмента x .

На обрабатываемом снимке контуры обнаруживаются как $m \times n$ -мерное изображение серого цвета. Для преобразования изображения в одномерную символьную строку направление отрезка, который, в свою очередь, содержит $w > 1$ элементов, представляется кодом одного из восьми изображенных на рис. 5.5 а) направлений.



а)



б)

Рис. 5.5. Принципы: а) кодирования направлений; б) выбора направления при разветвлении

Процедура поиска контуров при сканировании изображения заключается в нахождении элементов с ненулевой контрастностью. Поиск производится в направлении ориентации, которой соответствует наибольшее значение функции принадлежности. В случае же, когда максимальное значение функции принадлежности существует для двух и более направлений, как это изображено на рис. 5.5 б), используется ранее полученная информация и выбирается направление наибольшей связности, а впоследствии такой элемент рассматривается как развилка.

Помимо этого, при достижении определенной длины контура начало отрезка переносится, благодаря чему распознаются контуры в виде петель, и появляется возможность повторной обработки контуров в случае пропуска некоторого участка изображения.

На этапе сглаживания контуров используются следующие четыре процедуры:

1. Если подряд следуют четыре и более одинаковых кода, но между ними находится код другого направления, или два соседних кода, то этот код заменяется на последующий либо предыдущий, что изображено на рис. 5.6 а), либо исключается, что, в свою очередь, изображено на рис. 5.6 б);

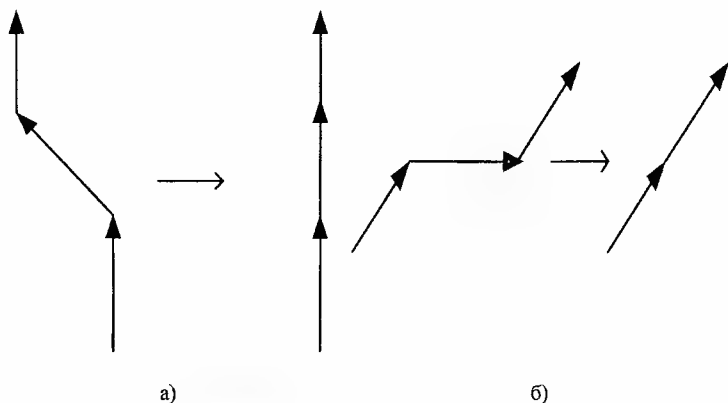


Рис. 5.6. Принципы: а) замещения кода; б) исключения кода

2. Если сначала следуют подряд четыре и более одинаковых кода, а затем два других кода, либо эти коды принимают значения 6 и 3, то, как и в предыдущем случае, эти коды заменяются на последующий либо предыдущий код, либо исключаются;
3. Если два соседних элемента имеют противоположные направления, то они исключаются;
4. Если некоторое направление незначительно изменяется, то для пары кодов в промежуточном направлении формируются правила вычисления векторной суммы, благодаря которым оказывается возможным устранить незначительное изменение направления.

Следующим этапом обработки изображения является сегментация контуров, которая представляет собой разграничение по изменениям свойств отрезков, в частности, граница выбирается в том месте, где изменение направления становится неодинаковым. Кривизна сег-

ментированных участков определяется в соответствии с формулой (5.4).

В качестве иллюстративного примера применения описанного подхода к обработке изображения приводится рентгеновский снимок запястья, представленный на рис 5.7, так что контуры костей представляют собой границы градиентов серого цвета для 145×128 элементов при $\gamma = 0.5$.



Рис. 5.7. Обрабатываемый рентгеновский снимок

Множество особенностей костей выбирается на основании вычисления значений принадлежностей линейности и кривизны сегментов. Полученный результат оказывается полезным при формировании правил распознавания неизвестных участков контуров. На рис. 5.8 представлен фрагмент исходного изображения а) до и б) после сглаживания, так что полученный результат позволяет достаточно легко выделять специфические особенности костей запястья пациента.

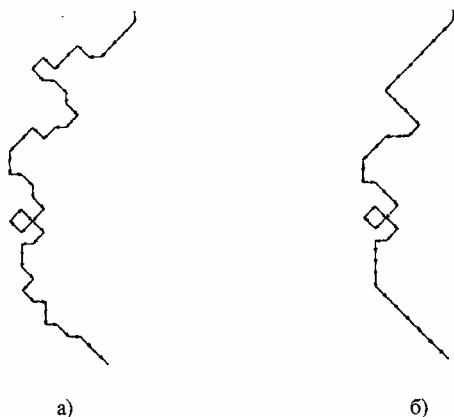


Рис. 5.8. Фрагмент обрабатываемого рентгеновского снимка: а) до сглаживания; б) после сглаживания

Различные нечеткие подходы к распознаванию образов предлагаются в процессе решения конкретных задач. Таким образом, их специфика зачастую оказывается обусловленной областью применения того или иного подхода. Обзор сфер применения нечетких методов распознавания образов приводится в таблице 5.5.

Таблица 5.5

Области применения нечетких методов распознавания образов

Область применения нечетких методов распознавания образов	Ссылка на источник
Распознавание речи	[75], [76], [135]
Интеллектуальные роботы	[100], [155]
Обработка изображений	[78], [102], [136], [138], [63], [41]
Распознавание символов	[124], [156], [157], [144]
Анализ сцен	[103], [104]
Распознавание геометрических объектов	[116], [165]
Классификация сигналов	[65], [66], [74], [141], [143]
Медицинские приложения	[89], [114], [153], [188]
Другие применения	[50], [70], [107]

Таковы основные подходы к решению задачи распознавания образов в условиях нестохастической неопределенности. В завершение следует указать, что в силу ограниченности изложения и значи-

тельному прогрессу в области нечетких методов распознавания образов данный обзор не является полным.

Резюме

Процесс определения нечеткой кластер-процедуры для решения задачи классификации в рамках прикладного исследования предусматривает два этапа: обоснование выбора одного из трех рассматриваемых типов методов нечеткого подхода к кластеризации и выбор конкретной кластер-процедуры. Рекомендации по выбору типа метода нечеткого подхода к решению задачи автоматической классификации можно выработать, исходя из целей классификации и имеющихся содержательных соображений о компактности выделяемых групп. Алгоритм нечеткого подхода решения задачи автоматической классификации, адекватный условиям конкретной задачи, может быть определен, исходя из вида матрицы исходных данных и имеющейся априорной информации. Если для решения задачи классификации выбрано эвристическое направление нечеткого подхода, то главным критерием выбора алгоритма является соответствие особенностей того или иного алгоритма содержательной постановке задачи. В случае, когда для решения задачи выбрано оптимизационное направление, главной проблемой оказывается обоснование вида функционала нечеткого разбиения. При рассмотрении этой проблемы должна учитываться, помимо формы матрицы исходных данных, вида шкалы, в которой измерены признаки, и типа признакового пространства, содержательная интерпретация функционала. Процедуры иерархического направления не предусматривают задания входных параметров, так что при выборе конкретной процедуры достаточно руководствоваться видом матрицы исходных данных и спецификой решаемой задачи. Ряд нечетких кластер-процедур предусматривает задание параметров работы. Число классов c зачастую может быть определено на основании сущности решаемой задачи. Если же число классов априори неизвестно, то целесообразно задать ряд значений c , $c_1 \leq c \leq c^*$. В алгоритме Кутюрье — Фьюлео число объектов в классе u , число объектов в области пересечения классов w и порог различия объектов α определяются спецификой решаемой задачи; в случае, когда жесткие ограничения на минимальное количество объектов в классе отсутствуют, целесообразно полагать $u=1$. Значение порога $\hat{d}$ выбирается в зависимости от используемой процедуры: в алгоритме Гитмана — Левина данный порог определяется исследователем в зависимости от вида функции принад-

лежности $\mu_B(x)$ в каждом конкретном случае, а в алгоритме Распини его можно задавать как среднюю связь в классе. Показатель нечеткости классификации γ чаще всего полагается равным 2, однако, в зависимости от особенностей конкретной задачи, может варьироваться в интервале от 1.5 до 30. Алгоритмы нечеткого подхода к решению задачи автоматической классификации находят самое разнообразное применение в задачах обработки изображений, распознавания рукописных символов, анализа сцен, классификации сигналов, в экологических, социально-экономических, медико-биологических исследованиях и при решении других задач. Среди других нечетких подходов к решению задач распознавания образов следует отметить нечеткий дискриминантный анализ, лингвистический подход, подход на основе упорядоченных взвешенных усредняющих операторов и подход на основе нечетких деревьев решений.

ЗАКЛЮЧЕНИЕ

Нечеткие методы автоматической классификации представляют собой мощный и гибкий аппарат проведения исследований и решения разнообразных задач практически в любой области человеческой деятельности. Главной особенностью нечетких методов классификации объектов в условиях отсутствия обучающих выборок является их сходство с человеческим способом классификации и, как следствие, простота интерпретации полученных результатов. Подытоживая рассмотрение современного состояния нечеткого подхода к решению задачи автоматической классификации, необходимо отметить две наблюдаемые тенденции его развития. Первое, на что следует обратить внимание — дальнейшее исследование и совершенствование существующих как собственно нечетких методов кластерного анализа, так и методов интерпретации и анализа полученных результатов. Второй тенденцией является создание новых нечетких методов распознавания образов с самообучением. Очевидно, что в ближайшие годы следует ожидать появления не только новых кластер-процедур как в рамках каждого из трех основных рассмотренных направлений, так и в рамках разработки аппроксимационного направления нечеткого подхода, но и новых подходов к разработке нечетких кластер-процедур. Представляется также значительное увеличение областей применения нечетких методов кластеризации: от медицинской диагностики на начальных стадиях развития заболевания до социально-политических исследований. Наиболее перспективным применение нечетких методов автоматической классификации представляется в военной области: от совершенствования существующих средств космической разведки до разработки принципиально новых средств обнаружения, идентификации, захвата, сопровождения и оценки поражения цели, включая создание новых систем самонаведения управляемых авиационных бомб, ракет класса «воздух-воздух» и класса «воздух-поверхность», совершенствования оптико-локационных систем для средств ПВО, а также при разработке иных образцов высокоточного оружия. Дальнейшее развитие нечетких методов автоматической классификации и расширение сфер их применения неизбежно повлечет увеличение количества работ в таких областях, как логика и методология научного познания и философия науки и техники.

Основные обозначения

$X = \{x_1, \dots, x_n\}$ — исследуемая совокупность объектов

$I^m(X)$ — пространство признаков

n — число объектов исследуемой совокупности

$\text{card}(X)$ — мощность множества X

c — число классов

$\vee$ — операция $\max$

$\wedge$ — операция $\min$

A_α — α -срез нечеткого множества A

$A_{(\alpha)}$ — нечеткое множество уровня α

$\mu_A(x)$ — функция принадлежности нечеткого множества A

A^l — l -е нечеткое множество; l -й нечеткий класс

μ_{li} — степень принадлежности i -го элемента $x_i \in X$ нечеткому классу A^l

$d(x_i, x_j)$ — расстояние между точками x_i и x_j

τ^l — типичная точка (центр, прототип) нечеткого класса A^l

$d(x_i, \tau^l)$ — расстояние от точки x_i до l -го прототипа τ^l

$P = \{A^1, \dots, A^c\}$ — нечеткое разбиение на c классов

$P_{\text{сзн}} = [\mu_{li}]$ — матрица размерности $(c \times n)$ нечеткого разбиения P

$F_c(P)$ — коэффициент нечеткого разбиения P

$H_c(P)$ — энтропия нечеткого разбиения P

$E_c(P)$ — пропорциональная экспонента нечеткого разбиения P

$X \times Y$ — декартово произведение множеств X и Y

R — произвольное бинарное нечеткое отношение на множестве X

$\mu_R(x_i, x_j)$ — функция принадлежности нечеткого отношения R

$R \circ S$ — $(\max\text{-}\min)$ -композиция нечетких отношений R и S

$\bar{R}$ — $(\max\text{-}\min)$ -транзитивное замыкание нечеткого отношения R

$\hat{R}$ — $(\min\text{-}\max)$ -транзитивное замыкание нечеткого отношения R

T — нечеткая толерантность (нечеткое отношение сходства) на множестве X

I — нечеткая интолерантность (нечеткое отношение несходства) на множестве X

S — нечеткая эквивалентность (отношение подобия) на множестве X

D — нечеткое отношение различия на множестве X

$X_{\text{сзн}} = [x'_i]$ — матрица размерности $(m \times n)$ значений признаков

$d_{\text{сзн}} = [d_{ij}]$ — матрица размерности $(n \times n)$ расстояний между объектами

$r_{\text{сзн}} = [r_{ij}]$ — матрица размерности $(n \times n)$ близостей между объектами

Список литературы

1. Аверкин А.Н. Нечеткое отношение моделирования и его использование для классификации и аппроксимации в нечетких лингвистических пространствах // Известия АН СССР. Техническая кибернетика. — 1982. — № 2. — с.215-217.
2. Аверкин А.Н., Тарасов В.Б. Нечеткое отношение моделирования и его применение в психологии и искусственном интеллекте. — М.: Вычислительный центр АН СССР, 1986. — 36с
3. Батыршин И.З. Кластеризация на основе размытых отношений сходства // Управление при наличии расплывчатых категорий: Тезисы докладов 3-го научно-технического семинара. — Пермь, 1980. — с.25-27.
4. Батыршин И.З., Вагин В.Н. Алгоритмы кластеризации, основывающиеся на понятии неразличимости объектов // Управление при наличии расплывчатых категорий: Тезисы докладов 4-го научно-технического семинара. — Фрунзе, 1981. — с.79.
5. Батыршин И.З. Иерархическая кластеризация на основе нечисловой информации о близости // Нечисловая статистика, экспертные оценки и смежные вопросы: Тезисы докладов II Всесоюзной конференции по статистическому и дискретному анализу нечисловой информации и экспертным оценкам. — Таллинн, 1984. — с.277.
6. Батыршин И.З. Иерархические алгоритмы выделения классов толерантности в задачах классификации // Применение вероятностно-статистических методов в бурении и нефтедобыче: Тезисы докладов IV Всесоюзной конференции. — Баку, 1984. — с.16-17.
7. Батыршин И.З., Халитов Р.Г. Иерархические алгоритмы кластеризации на базе классов толерантности // Повышение эффективности технологических процессов химических, нефтехимических и биотехнологических производств: Тезисы докладов Республиканской научно-практической конференции. — Казань: КХТИ, 1986. — с.109.
8. Батыршин И.З., Халитов Р.Г. Иерархическая классификация на базе классов толерантности // Исследование операций и аналитическое проектирование в технике. — Казань: КАИ, 1987. — с.105-110.
9. Батыршин И.З. О декомпозиции нечетких отношений эквивалентности // Математические и экспериментальные методы синтеза технических систем. — Казань: КАИ, 1989. — с.21-27.
10. Батыршин И.З., Морозов В.А., Халитов Р.Г. КЛАСТИЕР — программная система иерархической классификации // Статистический и дискретный анализ данных и экспертное оценивание: Материалы IV Всесоюзной школы-семинара. — Одесса, 1991. — с.319-321.
11. Берштейн Л.С., Дзюба Т.А. Решение задач классификации на нечетких графах // Новости искусственного интеллекта. — 2000. — № 3. — с. 113-121.
12. Блишун А.Ф. Отношение сходства нечетких понятий-классов // Управление при наличии расплывчатых категорий // Тезисы докладов 4-го научно-технического семинара. — Фрунзе, 1981. — с.79.
13. Бородин Л.И. Алгоритм построения линейной разделяющей функции с использованием нечетких разбиений // Компьютерный анализ данных и модели-

- рование: Сборник научных статей V Международной конференции (8-12 июня 1998 года, Минск). Часть 3: А-К / Под ред. проф. С.А. Айвазяна и проф. Ю.С. Харина. — Мн.: БГУ, 1998. — с.69-76.
14. Вятченин Д.А. Формы проявления нечеткости // Гуманитарно-экономический вестник. — 1998. — № 1. — с.66-69.
 15. Вятченин Д.А. Гносеологические аспекты применения теории нечетких множеств в кластер-анализе // «Великие преобразователи естествознания: Исаак Ньютон»: Тезисы докладов XVI Международных чтений. 29-30 ноября 2000 г., Минск / БГУИР. — Минск, 2000. — с.130-133.
 16. Вятченин Д.А. Содержательная интерпретация нечетких отношений сходства. // Полигнозис. — 2001. — № 1. — с. 20-25.
 17. Вятченин Д.А. О пересечении нечетких кластеров // Интегрированные модели и мягкие вычисления в искусственном интеллекте: Сборник трудов Международного научно-практического семинара, Коломна, 17-18 мая 2001. — М.: Наука, Физматлит, 2001. — с.122-126.
 18. Вятченин Д.А. Общая схема выбора типа метода и алгоритма нечеткого подхода в кластерном анализе // Новые информационные технологии: Материалы V Международной научной конференции НИТэ'2002 (Минск, 29-31 октября 2002 г.). Т.1. / Под ред. А.Н. Морозевича, Н.Н. Говядиновой, А.М. Зеневич, Л.С. Черепицы. — Минск: БГЭУ, 2002. — с.102-107.
 19. Вятченин Д.А. Основные концепции неопределенности в задачах автоматической классификации // Полигнозис. — 2002. — № 3. — с. 160-167.
 20. Давидовская И.А., Kovalik S. Применение нечеткого дерева решений для задач классификации // Новые информационные технологии: Материалы V Международной Научной Конференции НИТэ'2002 (Минск, 29-31 октября 2002 г.). Т.1. / Под ред. А.Н. Морозевича, Н.Н. Говядиновой, А.М. Зеневич, Л.С. Черепицы. — Минск: БГЭУ, 2002. — с.179-183.
 21. Данлюп С. Азбука звездного неба / Пер. с англ. — М.: Мир, 1990. — 240 с.
 22. Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов (Статистические методы классификации и измерения связей) — М.: Статистика, 1977. — 144 с.
 23. Жук Е.Е., Харин Ю.С. Устойчивость в кластер-анализе многомерных наблюдений. — Мн.: Белгосуниверситет, 1998. — 240 с.
 24. Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений / Пер. с англ.; под ред. Н.Н.Моисеева и С.А.Орловского. — М: Мир, 1976. — 165 с.
 25. Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер / Под ред. Дж. Вэн Райзина; пер с англ.; под ред. Ю.И.Журавлева. — М: Мир, 1980. — с. 208-247.
 26. Каркищенко А.Н., Бутенков С.А., Кривша В.В. Нечеткие геометрические признаки в задачах классификации и кластеризации // Новости искусственного интеллекта. — 2000. — № 3. — с. 129-133.
 27. Кофман А. Введение в теорию нечетких множеств / Пер. с фр.; Под ред. С.И. Травкина. — М.: Радио и связь, 1982. — 432 с.

28. Кузьмин В.Б. Построение групповых решений в пространствах четких и нечетких бинарных отношений. — М.: Наука, Гл. ред. физ.-мат. лит., 1982. — 168 с.
29. Лейбкинд А.Р., Рудник Б.Л., Тихомиров А.А. Математические методы и модели формирования организационных структур управления. — М.: МГУ, 1982. — 232 с.
30. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками / Пер. с англ. — М.: Финансы и статистика, 1990. — 336 с.
31. Мандель И.Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.
32. Миленский А. В. Классификация сигналов в условиях неопределенности. — М.: Сов. радио, 1975. — 328 с.
33. Нгуен М.Х. Применение нечетких отношений в классификации // Нечеткие системы поддержки принятия решений: Сборник научных трудов. — Калинин: Калининский государственный университет, 1989. — с. 99-107.
34. Нечеткие множества в моделях управления и искусственного интеллекта / А.Н.Аверкин, И.З.Батыршин, А.Ф.Блишун и др.; под ред. Д.А.Поспелова. — М.: Наука, гл. ред. физ.-мат. лит., 1986. — 312с.
35. Обработка нечеткой информации в системах принятия решений / А.Н. Борисов, А.В. Алексеев, Г.В. Меркурьева и др. — М.: Радио и связь, 1989. — 304с.
36. Ожегов С.И. Словарь русского языка: Ок. 57000 слов / Под ред. Н.Ю.Шведовой. — 17-е изд., стереотип. — М.: Русский язык, 1985. — 797с.
37. Прикладная статистика: Классификация и снижение размерности: Справ. изд./ С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; Под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1989. — 607 с.
38. Прикладные нечеткие системы: Пер. с япон. / К.Асаи, Д.Ватада, С.Иваи и др.; под ред. Т.Тэрано, К.Асаи, М.Сугэно. — М.: Мир, 1993. — 368с.
39. Рупини Э.Г. Последние достижения в нечетком кластер-анализе // Нечеткие множества и теория возможностей: Последние достижения / Под ред. Рональда Р. Ягера; Пер. с англ. В.Б. Кузьмина; Под ред. С.И. Травкина. — М.: Радио и связь, 1986. — с. 114-132.
40. Смоляк С.А., Титаренко Б.Н. Устойчивые методы оценивания. — М.: Статистика, 1982. — 208 с.
41. Тарасов В.Б., Желтов С.Ю., Степанов А.А. Нечеткие модели в обработке изображений: обзор зарубежных достижений // Новости искусственного интеллекта. — 1993. — № 3. — с.40-64.
42. Фукунага К. Введение в статистическую теорию распознавания образов / Пер. с англ. — М.: Наука, 1979. — 368 с.
43. Цихош Э. Сверхзвуковые самолеты: Справочное руководство / Пер. с польск.; под ред. В.Г. Микеладзе и Е.В. Зябрева. — М.: Мир, 1983. — 432 с.
44. Шлезингер М.И. О самопроизвольном различении образов // Читающие автоматы. — Киев: Наукова думка 1965. — с. 38-65.
45. Anderson E. The Irises of the Gaspe Peninsula // Bulletin of the American Iris Society. — 1935. — Vol. 59. — pp.2-5.
46. Backer E. Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets. Delft: Delft University Press; 1978.

47. Batyrshin I. Errors of Type 2 In Cluster Analysis and Invariant Cluster Procedures Based on similarity relations // Application of Fuzzy Systems: Proceedings of International Conference ICAFS-94 / Ed. by R. Aliev and R. Kenarangui. — University Press of Tabriz, Iran, 1994. — pp.374-378.
48. Bellacicco A. Fuzzy Classification // Synthese.— 1976.— Vol. 33.— pp.273-281.
49. Bellman R., Kalaba R., Zadeh L.A. Abstraction and Pattern Classification // Journal of Mathematical Analysis and Applications.— 1966.— Vol.13.— pp.1-7.
50. Behr D., Kocher R., Strackeljan J. Fuzzy Pattern Recognition for Automatic Detection of Different Teeth Substances // Fuzzy Sets and Systems.— 1997.— Vol.85.— pp.275-286.
51. Bezdek J.C. Fuzzy Mathematics in Pattern Classification. Ph.D. Thesis. Ithaca: Cornell University; 1973.
52. Bezdek J.C. Cluster Validity with Fuzzy Sets // Journal of Cybernetics.— 1974.— Vol. 3.— pp.58-73.
53. Bezdek J.C. Numerical Taxonomy With Fuzzy Sets // Journal of Mathematical Biology.— 1974.— Vol.1.— pp.57-71.
54. Bezdek J.C., Dunn J.C. Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions // IEEE Transactions on Computers.— 1975.— Vol. C-24.— pp.835-838.
55. Bezdek J.C. A Physical Interpretation of Fuzzy ISODATA // IEEE Transactions on Systems, Man, and Cybernetics.— 1976.— Vol. SMC-6.— pp.387-389.
56. Bezdek J.C., Castelaz P.F. Prototype Classification and Feature Selection with Fuzzy Sets // IEEE Transactions on Systems, Man, and Cybernetics.— 1977.— Vol. SMC-7.— pp.87-92.
57. Bezdek J.C., Harris J.D. Fuzzy Partitions and Relations: An Axiomatic Basis for Clustering // Fuzzy Sets and Systems.— 1978.— Vol.1.— pp.111-127.
58. Bezdek J.C., Windham M.P., Ehrlich R. Statistical Parameters of Cluster Validity Functionals // International Journal of Computer Information Sciences.— 1980.— Vol. 9.— pp.323-336.
59. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press; 1981.
60. Bezdek J.C., Coray C., Gunderson R., Watson J. Detection and Characterization of Cluster Substructure. Part I. Linear Structure: Fuzzy c-Lines; Part II. Fuzzy c-Varieties and Convex Combinations Thereof // SIAM Journal of Applied Mathematics.— 1981.— Vol. 40.— pp.339-372.
61. Bezdek J.C., Hathaway R.J., Howard R.E., Wilson C.A. Coordinate Descent and Clustering // Control and Cybernetics.— 1986.— Vol. 15.— pp.195-204.
62. Bezdek J.C., Tsao E.C.-K., Pal N.R. Fuzzy Kohonen Clustering Networks // Proceedings of the IEEE International Conference on Fuzzy Systems.— San Diego, 1992.— pp.1035-1043.
63. Bezdek J.C., Pal S.K. Fuzzy Models for Pattern Recognition. New York: IEEE Press; 1992.
64. Bocklisch S.F. Prozeßanalyse mit Unscharfen Verfahren. Berlin: Verlag Technik; 1987.

65. Bortolan G., Degani R. Ranking of Fuzzy Alternatives in Electrocardiography // Fuzzy Information, Knowledge Representation and Decision Analysis / Ed. by E. Sanchez and M.M. Gupta.— Oxford: Pergamon Press, 1983.— pp. 397-402.
66. Bortolan G., Degani R., Hirota K., Pedrycz W. Classification of ECG Signals — Utilization of Fuzzy Pattern Matching // Proceedings of International Workshop on Fuzzy Systems Application.— Iizuka; 1988.— p.88.
67. Chakrabarty M.K., Das M. On Fuzzy Equivalence. Part 1 // Fuzzy Sets and Systems.— 1983.— Vol.11.— pp.185-194.
68. Chakrabarty M.K., Das M. On Fuzzy Equivalence. Part 2 // Fuzzy Sets and Systems.— 1983.— Vol.11.— pp.299-308.
69. Chang R.L.P., Pavlidis T. Fuzzy Decision-Tree Algorithms // IEEE Transactions on Systems, Man, and Cybernetics.— 1977.— Vol. SMC-7.— pp.28-35.
70. Cheushev V. The Fuzzy Decisions and Contextual Support in Text Recognition by Information Theory Point of View // Pattern Recognition and Information Processing: Proceedings of Fourth International Conference (20-22 May 1997 Minsk, Republic of Belarus). Vol. 2. /Ed. by V. Krasnoproshin et al.— Szczecin: Wydawnictwo Uczelniane Politechniki Szczecinskiej, 1997.— pp.237-241.
71. Couturier A., Fioieau B. Recognising Stable Corporate Groups: A Fuzzy Classification Method // Fuzzy Economic Review.-1997.— Vol. II.— pp.35-45.
72. DataEngine Tutorials and Theory. Aachen: MIT GmbH; 1999.
73. Davé R.N., Sen S. Robust Fuzzy Clustering of Relational Data // IEEE Transactions on Fuzzy Systems.— 2002.— Vol. 10.— pp.713-727.
74. Degani R., Bortolan G. Fuzzy Numbers in Computerized Electrocardiography // Fuzzy Sets and Systems.— 1987.— Vol. 24.— pp.345-362.
75. Di Mori R., Laface P. Use of Fuzzy Algorithms for Phonetic and Phonenic Labeling of Continuous Speech // IEEE Transactions of Pattern Analysis Machines Intelligent.— 1980.— Vol. PAMI-2.— pp.136-148.
76. Di Mori R. Computerized Models of Speech Using Fuzzy Algorithms. New York: Plenum Press; 1983.
77. Dubes R.C., Jain A.K. Clustering Techniques: The User's Dilemma // Pattern Recognition.— 1976.— Vol. 8.— pp.247-260.
78. Dubois, D.; Jaulent, M.C. Some Techniques for Extracting Fuzzy Regions // Proceedings of First IFSA Congress (July 1-6, 1985, Mallorca, Spain). Vol. II / Mallorca, 1985.— pp.112-128.
79. Duda R.O., Hart P.E. Pattern Classification and Scene Analysis. New York: John Wiley & Sons; 1974.
80. Duin R.P.W. The Use of Continuous Variables for Labeling Objects // Pattern Recognition Letters.— 1982.— Vol.1.— pp.15-20.
81. Dumitrescu D. Hierarchical Pattern Classification // Fuzzy Sets and Systems.— 1988— Vol.28.— pp.145-162.
82. Dunn J.C. A Graph Theoretic Analysis of Pattern Classification via Tamura's Fuzzy Relation // IEEE Transactions on Systems, Man, and Cybernetics.— 1974.— Vol. SMC-4.— pp.310-313.
83. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics.— 1974.— Vol.3.— pp.32-57.

84. Dunn J.C. Well-separated Clusters and the Optimal Fuzzy Partitions // *Journal of Cybernetics*. — 1974.— Vol.4.— pp.95-104.
85. Dunn J.C. Some Recent Investigations of a New Fuzzy Partitioning Algorithm and Its Application to Pattern Classification Problems // *Journal of Cybernetics*. — 1974.— Vol.4.— pp.1-15.
86. Dunn J.C. Canonical Forms of Tamura's Fuzzy Relation: A Scheme for Visualizing Cluster Hierarchies // *Proceedings of International Conference on Computer Graphics, Pattern Recognition and Data Structure (May 1975)*. — Beverly Hills, 1975.
87. Freksa Ch. Linguistic Pattern Characterization and Analysis. Ph.D. Thesis. Berkeley: University of California; 1981.
88. Fu K.S. Syntactic Methods in Pattern Recognition. New York: Academic Press; 1984.
89. Gesn D.V., Maccarone M.C. Feature Selection and Possibility Theory // *Pattern Recognition*. — 1986.— Vol. 19.— pp.63-72.
90. Gitman J., Levine M.D. An Algorithm for Detecting Unimodal Fuzzy Sets and Its Application as a Clustering Technique // *IEEE Transactions on Computers*. — 1970.— Vol. C-19.— pp.583-593.
91. Gitman J. An Algorithm for Nonsupervised Pattern Classification // *IEEE Transactions on Systems, Man, and Cybernetics*. — 1973.— Vol. SMC-3.— pp.66-74.
92. Gustafson D.E., Kessel W.C. Fuzzy Clustering With a Fuzzy Covariance Matrix // *Advances in Fuzzy Set Theory and Applications* / Ed. by M.M.Gupta, R.K. Ragade, R.R. Yager. — Amsterdam: North-Holland Publishing Company, 1979.— pp.605-620.
93. Hartigan J.A. Clustering Algorithms. New York: John Wiley & Sons; 1975.
94. Hirota K. Concepts of Probabilistic Sets // *Fuzzy Sets and Systems*. — 1981.— Vol.1.— pp.31-46.
95. Hirota K. The Bounded Variation Quality and Its Application to Feature Extraction // *Pattern Recognition*. — 1982.— Vol.2.— pp.93-101.
96. Hirota K. Ambiguity Based on the Concept of Subjective Entropy // *Fuzzy Information and Decision Processes* / Ed. by M.M.Gupta, E.Sanchez. — Amsterdam: North-Holland Publishing Company, 1982.— pp. 29-40.
97. Hirota K., Pedrycz W. Characterization of Fuzzy Clustering Algorithms in Terms of Entropy of Probabilistic Sets // *Pattern Recognition Letters*. — 1984.— Vol. 2.— pp. 213-216.
98. Hirota K., Pedrycz W. Subjective Entropy of Probabilistic Sets and Fuzzy Cluster Analysis // *IEEE Transactions on Systems, Man, and Cybernetics*. — 1986.— Vol. SMC-16.— pp.173-179.
99. Hirota K., Iwami K., Pedrycz W. FCM-AD (Fuzzy Cluster Means with Additional Data) and Its Application to Aerial Images // *Proceedings of II IFSA Congress*. Vol. II.— Tokyo, 1987.— pp.729-732.
100. Hirota K. Fuzzy Robot Vision and Fuzzy Controlled Robot // *NATO ASI CIM* / Ed. by I.B. Turksen. — Berlin: Springer-Verlag, 1988.
101. Höppner F., Klawonn F., Kruse R. Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse. Wiesbaden, Braunschweig: Vieweg Verlag; 1997.

102. Huntsberger T.L., Jacobs C.L., Cannon R.L. Iterative Fuzzy Image Segmentation // Pattern Recognition.— 1985.— Vol.18.— pp.131-138.
103. Huntsberger T.L., Rangarajan Ch., Jayaramurthy S.N. Representation of Uncertainty in Computer Vision Using Fuzzy Sets // IEEE Transactions on Computers.— 1986.— Vol.C-35.— pp.145-156.
104. Jain R. Applications of Fuzzy Sets for the Analysis of Complex Scenes // Advances of Fuzzy Set Theory and Applications / Ed. by M.M. Gupta. — Amsterdam: North-Holland Publishing Company, 1979. — pp. 577-587.
105. Jajuga K. L_1 -norm Based Fuzzy Clustering // Fuzzy Sets and Systems.-1993.- Vol.39. -pp.43-50.
106. Jajuga K. Optimization in Fuzzy Clustering // Control and Cybernetics.-1995.— Vol. 24.— pp.409-419.
107. Kandel A. Fuzzy Techniques in Pattern Recognition. New York: John Wiley & Sons; 1982.
108. Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons; 1990.
109. Kaymak U., Setnes M. Fuzzy Clustering With Volume Prototypes and Adaptive Cluster Merging // IEEE Transactions on Fuzzy Systems.— 2002.— Vol. 10.— pp.705-712.
110. Kharin Yu.S. Robustness in Statistical Pattern Recognition. Dordrecht: Kluwer Academic Publishers; 1996.
111. Klir G.J., Folger T. Fuzzy Sets, Uncertainty and Information. NJ: Prentice Hall; 1988.
112. Koczy L.T., Hajnal M. A New Fuzzy Calculus and Its Applications as a Pattern Recognition Technique // Modern Trends in Cybernetics and Systems / Ed by J. Rose, C. Bilciu.— Berlin: Springer-Verlag, 1977.— pp. 103-118.
113. Kotoh K., Hiramatsu K. A Representation of Pattern Classes Using the Fuzzy Sets // Systems Computers Controls.-1973.— Vol.1-8.— pp.275-282.
114. Kumar A. A Real-Time System For Pattern Recognition of Human Sleep Stages by Fuzzy Systems Analysis // Pattern Recognition.-1977.— Vol.9.— pp.43-46.
115. Kuncheva L.I. An Application of OWA operators to the Aggregation of Multiple Classification Decisions // The Ordered Weighted Averaging Operators. Theory and Applications / Ed. by R.R. Yager, J. Kacprzyk. — Boston: Kluwer Academic Publishers, 1997.— pp.330-343.
116. Lee E.T. Proximity Measures For the Classification of Geometric Figures // Journal of Cybernetics.— 1972.— Vol.2.— pp.43-59.
117. Lee E.T. Application of Fuzzy Languages to Pattern Recognition // Kybernetes.— 1977.— Vol.6.— pp.167-173.
118. Lee E.T. Fuzzy Tree Automata and Syntactic Pattern Recognition // IEEE Transactions of Pattern Analysis Machines Intelligent.— 1982.— Vol. PAMI-4.— pp.445-449.
119. Libert G., Roubens M. Non-metric Fuzzy Clustering Algorithms And Their Cluster Validity // Approximate Reasoning in Decision Analysis / Ed. by M.M.Gupta, E.Sanchez. — Amsterdam: North-Holland Publishing Company, 1982.— pp.417-425.

120. Libert G. Compactness and Number of Clusters // *Control and Cybernetics*.— 1986.— Vol. 15.— pp.205-212.
121. Lim Y.W., Lee S.U. On the Color Image Segmentation Algorithm Based on the Thresholding and the Fuzzy c-Means Techniques // *Pattern Recognition*.— 1990.— Vol.23.— pp.935-952.
122. Lopez de Mantaras R., Aguilar Martin J. Classification and Linguistic Characterization of Nondeterministic Data // *Pattern Recognition Letters*.— 1983.— Vol. 2.— pp.33-41.
123. Lou S.-P., Cheng W.-C., Chao L.-M. Two New Methods In Fuzzy Cluster // *Approximate Reasoning in Decision Analysis* / Ed. by M.M.Gupta, E.Sanchez. — Amsterdam: North-Holland Publishing Company, 1982.— pp.427-430.
124. Majumder D.K.D., Pal S.K. Fuzzy Mathematical Approach to Pattern Recognition. Delhi: Wiley Eastern Ltd.; 1985.
125. Michalski R.S. Variable-Valued Logic and Its Applications to Pattern Recognition and Machine Learning // *Computer Science and Multiple-Valued logic: Theory and Application* / Ed. by D.C. Rine. — Amsterdam: North-Holland Publishing Company, 1975.— pp.506-534.
126. Michalski R.S. Pattern Recognition as Rule-Guided Inductive Inference // *IEEE Transactions of Pattern Analysis Machines Intelligent*.— 1980.— Vol. PAMI-2.— pp.349-361.
127. Miyakoshi M., Shimbo M. Solutions of Composite Fuzzy Relational Equations with Triangular Norms // *Fuzzy Sets and Systems*.— 1985.— Vol.16.— pp.53-63.
128. Miyamoto S. Fuzzy Sets in Information and Retrieval and Cluster Analysis. Dordrecht: Kluwer Academic Publishers; 1990.
129. Miyamoto S., Agusta Y. An Efficient Algorithm for L_1 Fuzzy c-Means and Its Termination // *Control and Cybernetics*.—1995.— Vol. 24.— pp.421-436.
130. Nath A.K., Lee T.T. On the Design of a Classifier With Linguistic Variables As Inputs // *Fuzzy Sets and Systems*.— 1983.— Vol.11.— pp. 265-286.
131. Nath A.K., Liu S.W., Lee T.T. On Some Properties of a Linguistic Classifier // *Fuzzy Sets and Systems*.— 1985.— Vol.17.— pp. 297-311.
132. Owsinski J.W. On a New Naturally Indexed Quick Clustering Method with a Global Objective Function // *Applied Stochastic Models and Data Analysis*.— 1990.— Vol. 6.— pp.157-171.
133. Owsinski J.W., Zadrozny S. Ecological Site Classification: An Application of Clustering. Reply to a Problem Proposed by H. El-Shishiny at the Fourth International Symposium on Applied Stochastic Models and Data Analysis, Nancy, France, December 7-9, 1988 // *Applied Stochastic Models and Data Analysis*.— 1991.— Vol. 7.— pp.273-279.
134. Owsinski J.W. Clustering — Modelling, Capacities, Limits, Applications // *Control and Cybernetics*.—1995.— Vol. 24.— pp.391-397.
135. Pal S.K., Majumder D.D. Fuzzy Sets and Decision-Making Approaches in Vowel and Speaker Recognition // *IEEE Transactions on Systems, Man, and Cybernetics*.— 1977.— Vol. SMC-7.— pp.625-629.
136. Pal S.K., King R.A. On Edge Detection of X-ray Images Using Fuzzy Sets // *IEEE Transactions of Pattern Analysis Machines Intelligent*.— 1983.— Vol. PAMI-1.— pp.67-77.

137. Pal S.K., King R.A., Hashim A.H. Image Description and Primitive Extraction Using Fuzzy Sets // IEEE Transactions on Systems, Man, and Cybernetics.— 1983.— Vol.SMC-13.— pp.94-100.
138. Pal S.K., Rosenfeld A. Image Enhancement and Thresholding by Optimization of Fuzzy Compactness // Pattern Recognition Letters.— 1990.— Vol. 11.— pp.831-841 Pattern Recognition Letters.— 1985.— Vol. 3.— pp.303-308..
139. Pedrycz W. Classification in a Fuzzy Environment // Pattern Recognition Letters.— 1985.— Vol. 3.— pp.303-308.
140. Pedrycz W. Algorithms of Fuzzy Clustering with Partial Supervision // Pattern Recognition Letters.— 1985.— Vol. 3.— pp.13-20.
141. Pedrycz W. ECG Signal Classification with the Aid of Linguistic Classifier // Proceedings of the XIV International Conference on Medical and Biomedical Engineering, Espoo 11-16 August 1985.— Espoo, 1985.
142. Pedrycz W. Techniques of Supervised and Unsupervised Pattern Recognition with the Aid of Fuzzy Set Theory // Pattern Recognition in Practice / Ed by E.S. Gelsema, L.N. Kanal. — Amsterdam: North-Holland Publishing Company, 1986.— pp. 439-448.
143. Pedrycz W., Gacek A. Feature Selection for ECG Signal Classification with the Aid of Qualitative and Quantitative Criteria // Proceedings of the 7th Hungarian Conference on Biomedical Engineering, 16-18 September 1987.— Hungary: Esztergom, 1987.
144. Pedrycz W. Fuzzy Sets in Pattern Recognition: Methodology and Methods // Pattern Recognition.— 1990.— Vol.23.— pp.121-146.
145. Pedrycz W. Formation of Prototypes and Their Confidence Regions in Classification and Concept Formation Problems // Pattern Recognition Letters.— 1991.— Vol. 12.— pp.739-746.
146. Pienkowski A.E. Artificial Color Perception Using Fuzzy Techniques in Digital Image Processing. Koln: Verlag TUV Rheinland; 1989.
147. Radecki T. Level fuzzy sets // Journal of Cybernetics.-1977.— Vol.7.— pp.189-198.
148. Roubens M. Pattern Classification Problems and Fuzzy Sets // Fuzzy Sets and Systems.— 1978.— Vol.1.— pp.239-253.
149. Ruspini E.H. A New Approach to Clustering // Information and Control.— 1969.— Vol.15.— pp.22-32.
150. Ruspini E.H. Numerical Methods for Fuzzy Clustering // Information Sciences.— 1970.— Vol.2.— pp.319-350.
151. Ruspini E.H. Optimization in Sample Descriptions — Data Reduction and Pattern Recognition Using Fuzzy Clustering // IEEE Transactions on Systems, Man, and Cybernetics.— 1972.— Vol.SMC-2.— pp.541.
152. Ruspini E.H. New Experimental Results in Fuzzy Clustering // Information Sciences.— 1973.— Vol.6.— pp.273-284.
153. Saitta L., Torasso P. Fuzzy Characteristics of Coronary Disease // Fuzzy Sets and Systems.— 1981.— Vol.5.— pp.245-258.
154. Sanchez E. Resolution of Composite Fuzzy Relation Equations // Information and Control.— 1976.— Vol.30.— pp.38-48.

155. Seif A., Aguilar-Martin J. Multi-Group Classification Using Fuzzy Correlation // Fuzzy Sets and Systems.-1980.— Vol.3.— pp.109-122.
156. Shimura M. Applications of Fuzzy Set Theory to Pattern Recognition // Journal of AACE.— 1975.— Vol.19.— pp.243-248.
157. Siy P., Chen C.S. Fuzzy Logic For Handwritten Numerical Character Recognition // IEEE Transactions on Systems, Man, and Cybernetics.— 1974.— Vol.SMC-4.— pp.570-575.
158. Straube B. Some Remarks to Fuzzy Cluster Analysis // Fuzzy Sets Applications, Methodological Approaches and Results: Proceedings of the International Workshop Co-sponsored by IIASA and Convened in Co-operation with IAEA held on the Wartburg, Eisenach (GDR) at March 3-8, 1985/Ed. by St.Bocklisch et al.— Berlin: Akademie— Verlag, 1986.— pp.165-174.
159. Straube B. Fuzzy Cluster Analysis By Means of FDA // Fuzzy Sets Applications, Methodological Approaches and Results: Proceedings of the International Workshop Co-sponsored by IIASA and Convened in Co-operation with IAEA held on the Wartburg, Eisenach (GDR) at March 3-8, 1985/Ed. by St.Bocklisch et al.— Berlin: Akademie— Verlag, 1986.— pp.175-180.
160. Sugeno M. Theory of Fuzzy Integrals and Its Applications. Ph.D. Thesis. Tokyo: Institute of Technology; 1971.
161. Sugeno M. Constructing Fuzzy Measure and Grading Similarity of Patterns by Fuzzy Integrals // Transactions of SICE.— 1973.— Vol.9.— pp.359-367.
162. Tamburrini G., Termini S. Some Foundational Problems in the Formalization of Vagueness // Fuzzy Information and Decision Processes / Ed. by M.M.Gupta, E.Sanchez. — Amsterdam: North-Holland Publishing Company, 1982.— pp.161-166.
163. Tamura S., Higuchi S., Tanaka K. Pattern Classification Based on Fuzzy Relations // IEEE Transactions on Systems, Man, and Cybernetics.— 1971.— Vol.SMC-1.— pp.61-66.
164. Termini S. Aspects of Vagueness and Some Epistemological Problems Related to Their Formalization // Aspects of Vagueness/Ed. by H.J. Skala, S. Termini, E. Trillas.— Dordrecht: D.Reidel Publishing Company, 1984.— pp.205-230.
165. Tou J.T. An Approach to Understanding Geometrical Configurations by Computer // International Journal of Computer and Information Sciences.— 1980.— Vol.9., N 2.— pp.1-13.
166. Trauwaert E. On the Meaning of Dunn's Partition Coefficient for Fuzzy Clusters // Fuzzy Sets and Systems.— 1988.— Vol.25.— pp.217-242.
167. Trauwaert E., Kaufman L., Rousseeuw P. Fuzzy Clustering Algorithms Based on the Maximum Likelihood Principle // Fuzzy Sets and Systems.— 1991.— Vol.42.— pp.213-227.
168. Trivedi M.M., Bezdek J.C. Low-level Segmentation of Aerial Images with Fuzzy Clustering // IEEE Transactions on Systems, Man, and Cybernetics, 1986, vol. SMC-16, pp.589-596.
169. Trivedi M.M. Analysis of Aerial Image Using Fuzzy Clustering // Analysis of Fuzzy Information / Ed. by J.C. Bezdek. — CDC Press, 1987, Vol.3.— pp.133-151.
170. Vatlín S.I. Selfguessing Fuzzy Classifiers // Informatica.— 1993.— Vol.4.— pp.406-414.

171. Vatin S.I. The Covariant Monotonic Improvement of Fuzzy Classifiers is Impossible // *Neural Network World*.— 1996.— Vol.6.— pp.401-406.
172. Vatin S.I. The Structure of the Improvement of Consistently Fuzzy Classifiers // *Pattern Recognition and Information Processing: Proceedings of Fourth International Conference (20-22 May 1997 Minsk, Republic of Belarus)*. Vol. 2. /Ed. by V. Krasnoproshin et al. — Szczecin: Wydawnictwo Uczelniane Politechniki Szczecinskiej, 1997.— pp.87-90.
173. Viattchenin D.A. Fuzziness As A Theoretical Concept // *Proceedings of the European Symposium on Intelligent Techniques, March 20-21, 1997, Bari, Italy*.— Aachen: ERUDIT Service Center, 1997.— pp.258-260.
174. Viattchenin D.A. Some Remarks To Concept of Fuzzy Similarity Relation For Fuzzy Cluster Analysis // *Pattern Recognition and Information Processing: Proceedings of Fourth International Conference (20-22 May 1997 Minsk, Republic of Belarus)*. Vol. 2. /Ed. by V. Krasnoproshin et al.— Szczecin: Wydawnictwo Uczelniane Politechniki Szczecinskiej, 1997.— pp.35-38.
175. Viattchenin D.A. On Projections of Fuzzy Similarity Relations // *Computer Data Analysis and Modeling: Proceedings of the Fifth International Conference (June 8-12, 1998, Minsk)*. Vol. 2: M-Z/Ed. By Prof. S.A.Aivazyan and Prof. Yu.S.Kharin. — Minsk, BSU, 1998. — pp. 150-155.
176. Viattchenin D.A. Application of Fuzzy Similarity Relation to Data Representation for Pattern Classification Problems // *Advanced Computer Systems: Proceedings of Sixth International Conference (Szczecin, Poland, November 1999)* /Ed. by J. Soldek, J. Pejas. — Szczecin: "INFORMA", 1999.— pp.26-28.
177. Viattchenin D.A. Theoretical Notes on Fuzzy Intolerances. // *Pattern Recognition and Information Processing: Proceedings of Sixth International Conference (15-17 May 2001, Minsk, Republic of Belarus)*. Vol. 2/Ed. by S. Ablameyko et al. — Minsk, Institute of Engineering Cybernetics of the National Academy of Sciences of Belarus, 2001. — pp. 142-146.
178. Viattchenin D.A. Human-Computer Approach to Fuzzy Classification Problem Based on the Concept of Representation // *Quality of Life: Proceedings of the Second International Conference (Wroclaw, 18-20 September 2002)* / Ed. by Walenty Ostasiewicz. — Wroclaw: Wroclaw University of Economics, 2002. — pp. 189-204.
179. Vila M.A., Delgado M. Problems of Classification in a Fuzzy Environment // *Fuzzy Sets and Systems*.— 1983.— Vol.3.— pp. 229-239.
180. Walesiak M. Uogólniona Miara Odległości w Statystycznej Analizie Wielowymiarowej. Wroclaw: Wydawnictwo AE; 2002.
181. Wang X., Chen B., Qian G., Ye F. On the Optimization of Fuzzy Decision Trees // *Fuzzy Sets and Systems*.— 2000.— Vol.112.— pp.117-125.
182. Watada J., Tanaka H., Asai K. A Heuristic Method of Hierarchical Clustering for Fuzzy Intransitive Relations // *Fuzzy Set and Possibility Theory* / Ed. by R.R.Yager. — New York: Pergamon Press, 1982.— pp.148-166.
183. Watanabe S. *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons; 1985.
184. Windham M.P. Cluster Validity for Fuzzy Clustering Algorithms // *Fuzzy Sets and Systems*.— 1981.— Vol.2.— pp. 177-186.

185. Windham M.P. Geometric Fuzzy Clustering Algorithms // Fuzzy Sets and Systems.— 1983.— Vol.3.— pp. 271-280.
186. Windham M.R. Numerical Classification of Proximity Data with Assignment Measures // Journal of Classification.— 1985.— Vol.2.— pp.157-172.
187. Wong M.A. A Hybrid Clustering Method for Identifying High-Density Clusters // Journal of American Statistical Association.— 1982.— Vol. 77.— pp.841-847.
188. Woodbury M.A., Clive J. Clinical Pure Types as a Fuzzy Partition // Journal of Cybernetics.— 1974.— Vol. 3. — pp.111-121.
189. Wright W. A Formalization of Cluster Analysis // Pattern Recognition Letters.— 1975.— Vol. 5.— pp.273-282.
190. Yeh R.T., Bang S.Y. Fuzzy Relations, Fuzzy Graphs and Their Applications to Clustering Analysis // Fuzzy Sets and Their Applications to Cognitive and Decision Processes / Ed. by L.A. Zadeh and al. — New York: Academic Press, 1975.— pp.125-149.
191. Yuan Y., Shaw M.J. Induction of Fuzzy Decision Trees // Fuzzy Sets and Systems.— 1995.— Vol.69.— pp.125-139.
192. Zadeh L.A. Fuzzy Sets // Information and Control.— 1965.— Vol.8.— pp.338-353.
193. Zadeh L.A. Similarity Relations and Fuzzy Orderings // Information Sciences.— 1971.— Vol.3.— pp.177-200.

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ	3
ГЛАВА 1	
ПРОБЛЕМА НЕОПРЕДЕЛЕННОСТИ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ.....	7
1.1. Понятие однородности и проблема классификации объектов.....	7
1.1.1. Основные подходы к решению проблемы выделения однородных групп объектов.....	7
1.1.2. Общая постановка задачи автоматической классификации и основные направления ее решения	11
1.2. Подход к решению задачи автоматической классификации с позиции теории нечетких множеств	17
1.2.1. Основные концепции неопределенности в задачах автоматической классификации.....	17
1.2.2. Гносеологические аспекты подхода к решению задач классификации с позиции теории нечетких множеств.....	29
Резюме.....	36
ГЛАВА 2	
ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ.....	39
2.1. Нечеткие множества	39
2.1.1. Понятие взвешенной принадлежности и основные определения теории нечетких множеств.....	39
2.1.2. Основные операции над нечеткими множествами	43
2.2. Нечеткие отношения.....	46
2.2.1. Определение нечеткого отношения; свойства и классификация нечетких отношений.....	46
2.2.2. Основные операции над нечеткими отношениями	51
Резюме.....	54
ГЛАВА 3	
МЕТОДЫ НЕЧЕТКОГО ПОДХОДА К РЕШЕНИЮ ЗАДАЧ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ	55

3.1. Эвристические методы	55
3.1.1. Особенности эвристических методов решения нечеткой модификации задачи автоматической классификации	55
3.1.2. Описание алгоритмов.....	57
3.1.2.1. Алгоритм Гитмана — Левина	57
3.1.2.2. Алгоритм Тамуры — Хигути — Танаки.....	62
3.1.2.3. Алгоритм Кутюрье — Фьолео	63
3.1.2.4. Алгоритм Берштейна — Дзюбы	66
3.1.2.5. Другие алгоритмы	69
3.2. Оптимизационные методы	70
3.2.1. Нечеткая модификация экстремальной постановки задачи автоматической классификации и функционалы качества нечеткого разбиения	70
3.2.2. Описание алгоритмов.....	94
3.2.2.1. Алгоритм Распини	94
3.2.2.2. Алгоритм Уиндхема	97
3.2.2.3. Алгоритм Рубенса	99
3.2.2.4. Алгоритм Беждека — Данна	100
3.2.2.5. Алгоритм Педрича	101
3.2.2.6. Алгоритм Даве — Сена.....	102
3.2.2.7. Другие алгоритмы	103
3.3. Иерархические методы	111
3.3.1. Основные понятия иерархического направления решения нечеткой модификации задачи автоматической классификации.....	111
3.3.2. Описание алгоритмов.....	112
3.3.2.1. Алгоритм Ватады — Танаки — Асаи.....	112
3.3.2.2. Алгоритм Думитреску	121
3.3.2.3. Другие алгоритмы	130
Резюме	132

ГЛАВА 4

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕЧЕТКИХ МЕТОДОВ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ	134
---	-----

4.1. Содержательные аспекты сравнения нечетких методов автоматической классификации	134
4.1.1. Цели сравнения и оценка сходимости нечетких кластер-процедур.....	134
4.1.2. Виды исходных данных и общая схема проведения сравнительного анализа.....	135
4.2. Оценка и представление результатов обработки данных нечеткими методами автоматической классификации	140
4.2.1. Оценка качества нечеткой классификации	140

4.2.2. Представление и интерпретация результатов нечеткой классификации.....	149
4.3. Экспериментальное сравнение нечетких методов автоматической классификации	151
4.3.1. Исходные данные для проведения экспериментального сравнения алгоритмов	151
4.3.2. Результаты вычислительных экспериментов	156
Резюме.....	176
 ГЛАВА 5	
 МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ НЕЧЕТКОГО ПОДХОДА К РЕШЕНИЮ ЗАДАЧИ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ.....	177
5.1. Общая схема применения нечетких методов решения задачи автоматической классификации.....	177
5.1.1. Выбор типа метода нечеткого подхода к решению задачи автоматической классификации.....	177
5.1.2. Выбор алгоритма нечеткого подхода к решению задачи автоматической классификации и обоснование параметров	179
5.2. Прикладные аспекты нечеткой классификации.....	185
5.2.1. Применение нечетких методов автоматической классификации к обработке изображений	185
5.2.2. Нечеткие методы автоматической классификации и другие нечеткие методы распознавания образов	189
Резюме.....	198
 ЗАКЛЮЧЕНИЕ	200
 ОСНОВНЫЕ ОБОЗНАЧЕНИЯ.....	201
 СПИСОК ЛИТЕРАТУРЫ	202

Вятченин Дмитрий Аркадьевич

**НЕЧЕТКИЕ МЕТОДЫ
АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ**

Монография

Ответственный за выпуск *А.П. Аношко*

Редактор *Г.В. Ширкина*

Технический редактор *О.А. Курятова*

Сдано в набор 15.11.03. Подписано в печать 10.05.04.

Бумага офсетная. Формат 60х84/16.

Гарнитура Таймс. Печать офсетная.

Усл. печ. л. 12,4. Уч. изд. л. 12,8.

Тираж 200 экз. Заказ 028.

Издано на УП «Технопринт», лицензия № 02330/0056932.

Отпечатано на УП «Технопринт», ЛП № 203 от 26.01.03

220027, Минск, пр-т Ф. Скорины, 65, корп. 14, ком. 205а.

Тел/факс 231-86-93